From Hooks to Clicks: A Data-driven Approach to Understanding Language Trends in Phishing Schemes Across Different Attack Vectors

by Luma M. Hawamdah

B.S. in Computer Science, December 2006, University of Jordan
M.S. in Information Technology, July 2015, Southern New Hampshire University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 5, 2024

Praxis directed by

Mehdi Hashemipour
Professorial Lecturer in Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Luma Mousa Hawamdah has passed the Final Examination for the degree of Doctor of Engineering as of December 01, 2023. This is the final and approved form of the Praxis.

From Hooks to Clicks: A Data-driven Approach to Understanding Language Trends in Phishing Schemes Across Different Attack Vectors

Luma M. Hawamdah

Praxis Research Committee:

Mehdi Hashemipour, Professorial Lecturer in Engineering and Applied Science, Praxis Director

John Sahlin, Professorial Lecturer in Engineering and Applied Science, Committee Member

Muhammad F. Islam, Professorial Lecturer in Engineering and Applied Science, Committee Member

© Copyright 2023 by Luma M. Hawamdah
All rights reserved

Dedication

I want to thank my daughter Layan, who without her support and love I would not have been able to keep going in this world. She is my power, motivation, and determination to continue growing and learning to be a strong role model for her. This work is for you, and all the sacrifices you've made with me along the way. I could not have done it without you.

Acknowledgments

First and foremost, thank you God for all the guidance and light you shed in my life. Working on this Doctorate has been a life-changing experience for me, and it wouldn't have been possible for me to go through it without the people below:

o suitable words convey my deep gratitude and respect for my family. My sister and brother, Lana and Souheib Hawamdah, have been my rocks through life and have always supported me and believed in me. Thank you for being you. My parents – Mom and Dad – thank you for always loving me and pushing me to keep going. Thank you to the friends I call family who continue to support me and my big dreams.

My sincere thanks must also go to the members of my praxis advisory especially research advisor, Mehdi Hashemipour, for his guidance and expertise throughout the praxis research process. Finally, my program mates who continued to give their time to offer valuable feedback and comments for me to improve my work and reach my goal.

Abstract of Praxis

From Hooks to Clicks: A Data-driven Approach to Understanding Language Trends in Phishing Schemes Across Different Attack Vectors

Phishing attacks utilize cybersecurity's significant vulnerabilities, human beings, by leveraging the skills of human deceptiveness and technology to turn their targets into victims. These attacks have become the number one preferred method for adversaries to execute their attacks by successfully understanding the victim's behavior and emotions to gain their trust. They pose a high threat to organizations in both the business and personal aspects of it. Even though organizations invest in security systems, and updates to their infrastructure they have yet to find a complete defense mechanism against these attacks where individuals follow secure protocols, security awareness training for organizations, and rules and regulations enacted to defend against phishing attacks.

With the above understanding in mind, a novel topic to investigate a different form of analysis was drawn out in the open. This research aims to address the critical challenge of identifying phishing attempts within various forms of digital communication, such as emails, text messages, malicious URLs, and job postings. The goal is to leverage Natural Language Processing (NLP) techniques, specifically Bidirectional Encoder Representations from Transformers (BERT), to create a robust model capable of flagging potential phishing content from a diverse range of datasets. The base model introduced is the "**PhishNet**". It was further decided to compare the same framework on BART and XLNET models to verify that the BERT approach with the concatenated dataset provides valid credibility to the approach.

The model was implemented by assessing a basic predictive model for detecting phishing URLS and building on it by utilizing multi-diverse datasets that were concatenated to address the gaps identified in the predictive models in today's market. The data was facilitated by different phishing attacks building upon it and combining the data into one form of entry to provide us with a larger scale of understanding of the language used to perform a phishing attack.

The phishing dataset shows that there are six main categories of language that are used to grab the attention of the victims – these include obtaining a reward, a form of urgency, curiosity, job requirement, for entertainment purposes and from fear. PhishNet has proven to provide higher accuracy than the current standalone URL blocklist model with an accuracy result of over 99% as it identifies these different psychological traits an attacker might have to perform the attack. With this understanding, we are also able to create a framework for training employees and investigating the feasibility of building the model through a browser extension. The research's significance lies in offering innovative and effective approaches to prevent and mitigate phishing assaults and strengthening information system security.

Keywords: social engineering attacks, phishing, predictive model, phishing attacks, user characteristics, data breaches, cybersecurity, information system security, natural language processing (NLP), machine learning model.

Table of Contents

Dedication ..... **iv**

Acknowledgments ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xii**

List of Appendix Figures ..... **xv**

List of Tables ..... **xvi**

— Introduction **1**

1.1 Background 1

1.2 Research Motivation 4

1.3 Problem Statement 6

1.4 Thesis Statement 7

1.5 Research Objectives 8

1.6 Research Questions and Hypotheses 9

1.7 Scope of Research 10

1.8 Research Limitations 11

1.9 Organization of Praxis 12

— Literature Review **14**

2.1 Introduction 14

2.2 Review of Social Engineering Attacks 15

2.2.1 Existing Research on Social Engineering Attacks 15

2.2.2 Types Social Engineering Attacks 17

2.2.3 How to Avoid or Minimize These Attacks 18
2.2.4 Techniques Used in Social Engineering Attacks 20
2.2.5 Techniques to Thwart Social Engineering Attacks 22
2.3 Detection of Social Engineering with Predictive Models 23
2.4 Phishing attacks 24
2.4.1 Types of Phishing Attacks 31
2.4.2 Detecting Phishing Attacks 32
2.4.3 Anatomy of Phishing Messages 34
2.4.4 Impact and Statistics of Phishing Attacks 39
2.5 Machine Learning-Based Methods 42
2.5.1 Current Phishing Detection Models 42
2.5.2 Phishing Detection Framework 45
2.5.3 Human and Organizational Error 47
2.6 User Behavior 49
2.7 Solution to Phishing Attacks 52
2.8 Summary and Conclusion 55
— Methodology 57
3.1 Introduction 57
3.2 Research Approach 60
3.3 Research Setting and Context 61
3.4 Data Collection Methodology 62
3.4.1 Dataset 1: Phishing GPT Message Dataset 62
3.4.2 Dataset 2: Phishing Text Messages Dataset 63

3.4.3 Dataset 3: Fake Job Posting Dataset....63
3.4.4 Dataset 4: Legit and Phishing Emails Dataset....63
3.4.5 Dataset 5: Phishing URLs Dataset....63
3.4.6 Concatenation of Datasets....64
3.5 Data Analysis Tool....65
3.6 Learning Algorithms Used....66
3.6.1 Natural Language Processing....68
3.6.2 Bidirectional Encoder Representations from Transformers....68
3.6.3 Bidirectional and Auto-Regressive Transformers....69
3.6.4 XLNet....69
3.7 Model Architecture....69
3.8 Development of a New Framework....72
3.9 Conclusion....74
— Results....76
4.1 Introduction....76
4.2 Dataset Quality Metrics....79
4.2.1 Data Feature Engineering....80
4.2.2 Data Pre-Processing and Exploratory Data Analysis....84
4.2.3 Identifying and Rectifying Data Quality Issues in Training Data....94
4.3 Data Assessment and Evaluation....95
4.3.1 Standalone Datasets Assessment....96
4.3.2 Standalone DataSets Evaluation on PhishNet Model....99
4.3.3 Concatenated Dataset Assessment....103

4.3.4 Concatenated Dataset Model Evaluation....103
4.4 PhishNet Model Inference....104
4.5 The PhishNet Model....110
4.6 Analysis of Model's Performance....111
4.6.1 Dataset 1: Phishing GPT Messages Dataset....112
4.6.2 Dataset 2: Phishing Text Messages Dataset....112
4.6.3 Dataset 3: Fake Job Posting Dataset....113
4.6.4 Dataset 4: Phishing Emails Dataset....114
4.6.5 Dataset 5: Malicious URL Dataset....115
4.6.6 Concatenated Dataset....116
4.7 BERT vs BART and XLNET....117
4.8 Hypothesis Testing....122
4.8.1 Hypothesis 1....122
4.8.2 Hypothesis 2....124
4.8.3 Hypothesis 3....125
4.9 Conclusion....126
— Discussion and Conclusion....129
5.1 Discussion....129
5.2 Contributions to Body of Knowledge....130
5.3 Future Recommendations....131
5.4 Conclusion....132
References....134
Appendices....145

List of Figures

Figure 1: The Social Engineering Personality Framework.....14

Figure 2: Overview of Verizon 2020 Data Breach Report. 16

Figure 3: Impact of Social Engineering Attacks.....17

Figure 4: Social Engineering Life Cycle 19

Figure 5: Phishing Process Flow and Phases.....27

Figure 6: Most Targeted Phishing Sectors.....31

Figure 7: Phishing Email Anatomy for Microsoft Office.....37

Figure 8: Phishing Email Example for Payroll Setting.....38

Figure 9: Phishing Email Example from the IRS 38

Figure 10: 2021 The Sources of Phishing Attacks.....40

Figure 11: Industries Targeted by Phishing Attacks.....42

Figure 12: One Approach Machine Learning for Phishing Attack Detection 43

Figure 13: The Analysis Framework for Phishing Attacks 46

Figure 14: Detecting Website 50

Figure 15: Solution to Phishing 54

Figure 16: Praxis Research Methodology.....60

Figure 17: Data Distribution of the Six Features Present in Phishing Datasets.....81

Figure 18: Correlation Heatmap of the Six Features 82

Figure 19: Decision Tree vs Random Forest Classification of Phishing Data Features....83

Figure 20: Exploratory Data Analysis for Dataset-1- Phishing GPT Messages.....86

Figure 21: Exploratory Data Analysis for Dataset-2- Phishing Text Messages 87

Figure 22: Exploratory Data Analysis for Dataset-3: Fake Job Posting.....88

Figure 23: Exploratory Data Analysis for Dataset-4 – Phishing Emails 90

Figure 24: Exploratory Data Analysis for Dataset-5 – Malicious URL 91

Figure 25: Exploratory Data Analysis for Concatenation of Datasets.....94

Figure 26: Data Validation Process 96

Figure 27: Phishing Text Accuracy and Precision of its Categorization 98

Figure 28: Distribution of Phishing and Legit Email Length 99

Figure 29: Training and Evaluation of Dataset 1: Phishing GPT Message 100

Figure 30: Training and Evaluation of Dataset 2: Phishing Text Message 101

Figure 31: Training and Evaluation of Dataset 3: Fake Job Posting 101

Figure 32: Training and Evaluation of Dataset 4: Phishing Emails.....102

Figure 33: Training and Evaluation of Dataset 5: Malicious URL.....102

Figure 34: Training and Evaluation of Concatenation of all Datasets.....104

Figure 35: Heatmap of Test Dataset 1: Phishing GPT Message.....105

Figure 36: Heatmap of Test Dataset 2: Phishing Text Message.....106

Figure 37: Heatmap of Test Dataset 3: Fake Job Posting.....107

Figure 38: Heatmap of Test Dataset 4: Phishing Email.....108

Figure 39: Heatmap of Test Dataset 5: Malicious URL 109

Figure 40: Heatmap of Test of Concatenation of all Datasets 110

Figure 41: Snippet of the BART Model 119

Figure 42: Snippet of the XLNET Model 120

Figure 43: Training and Evaluation of Concatenation of all Datasets on BART Model.121

Figure 44: Training and Evaluation of Concatenation of all Datasets on XLNET Model 121

List of Appendix Figures

Figure A-1: Phishing Email Dataset Analysis on Multinominal 145

Figure A-2: Phishing Email Dataset Analysis on Random Forest.....145

Figure A-3: Phishing Text Message Number of Character Distribution 146

Figure A-4: Phishing Text Message Number of Words Distribution.....146

Figure A-5: Phishing Text Message Feature Relationship 147

Figure A-6: Common Phishing Words in Phishing Text Dataset.....147

Figure A-7: Common Legit Words in Phishing Text Dataset 148

Figure A-8: URL Feature Engineering Distribution.....148

Figure A-9: URL Feature Engineering Feature Importance Distribution.....149

List of Tables

Table 1: Comparison of Machine Learning Based on Phishing Detection Systems 34

Table 2: Model Architecture.....71

Table 3: Phishing Security Controls Developed Through the MITRE Attack Framework 74

Table 4: Average of Each Feature in Phishing Data.....82

Table 5: Compare the Models Performance on Feature Engineering.....84

Table 6: Descriptive Analysis of Phishing GPT Message Dataset.....86

Table 7: Descriptive Analysis of Phishing Text Message Dataset 87

Table 8: Descriptive Analysis of: Fake Job Posting Dataset 89

Table 9: Descriptive Analysis of: Phishing Emails Dataset 90

Table 10: Descriptive Analysis of: Malicious URL Dataset 92

Table 11: Confusion Matrix for Test Dataset 1: Phishing GPT Message.....105

Table 12: Confusion Matrix for Test Dataset 2: Phishing Text Message.....106

Table 13: Confusion Matrix for Test Dataset 3: Fake Job Posting.....107

Table 14: Confusion Matrix for Test Dataset 4: Phishing Email.....108

Table 15: Confusion Matrix for Test Dataset 5: Malicious URL 109

Table 16: Confusion Matrix for Test of Concatenation of all Datasets.....110

Table 17:Comparasion of BART and XLNET Models 1108

— Introduction

1.1 Background

Confidential data is highly used in both personal and business settings. The internet has become a massive space where people from all over the world communicate with one another and through this form of communication, many of the data exchanges done includes sensitive data. This transfer of data makes the communication vulnerable to cyber-attacks that later uses the data for fraudulent activities.

Compromise of both personal and business confidential information through social engineering is a growing source of information security concern. The term *social engineering* is used for a selection of widespread adversary actions achieved through interactions with humans through deceiving them. This type of enemy assault has been happening since the Greek times. The Greeks attempted to slip into the city of Troy with a diversion to sidestep the line of safeguard at the doors and assault the city (Rees, 2022). Like the Greeks, adversaries use social engineering techniques to manipulate humans into handing over sensitive information on bases of trusting them. It uses mental and emotional manipulation to deceive victims into compromising the security measures in place and providing sensitive information to the attacker using tactics that could potentially be used for fraudulent behavior (Salahdine & Kaabouch, 2019).

Social engineering attacks happen in multiple attempts. The attacks can be accomplished through baiting, scareware, phishing, pretexting, and spear phishing (Uebelacker et al., 2014). A cybercriminal first studies and examines the intended victim and collects all necessary information about their background, from security protocols to entry points needed to perform the attack. Afterwards the cybercriminal performing the

attack takes on the action of gaining the victim's trust and provide incentives for succeeding actions that disrupt security practices, which include revealing classified information or allowing access to critical resources (Uebelacker et al., 2014).Kevin Mitnick stated on BBC news: "The biggest threat to the security of a company is not a computer virus, an unpatched hole in a key program or a badly installed firewall. In fact, the biggest threat could be **you**. What I found personally to be true was that it's easier to manipulate people rather than technology. Most of the time organizations overlook that human element." (Kim, 2023).

Alzahrani (2020) states that social engineering attacks rely on manipulating human psychology rather than technical exploits. These attacks, which frequently target human error and trust, can have severe consequences, including data breaches, financial losses, or identity theft (Zhang et al., 2022). In correspondence to the Verizon Data Breach Investigations Report from 2020, social engineering was implicated in over 20% of data breaches, making it the second most common cause of such breaches after hacking (Nathan & Scobell, 2021). Furthermore, SOCRadar gave an overview of the breach report 2020 and determined that social engineering dominated by phishing attacks was 22% (SOCRadar, 2020). Phishing attacks account for 90% of data breaches and 84% of organizations were the targets of at least one phishing attempt in 2023.

Phishing websites, emails, social media messages, and texts are forged messages with links that are generated by malicious individuals looking to mimic webpages of real websites (Alkhalil, et al., 2021). Clicking on a defected link that was maliciously created could lead to the installation of a malicious form of software and open the gateway for adversaries to steal sensitive information or freeze a system to become part of a

ransomware attack. Victims of these forms of attacks may also disclose their financial information to the criminal who will use the gained information for financial and criminal activities (SOCRadar, 2020).

Effective strategies for preventing and detecting social engineering attacks, such as phishing, involves using predictive models to identify trends and user characteristics associated with these attacks. Using predictive models to analyze past attacks can identify attackers tactics and victims' demographics (SOCRadar, 2020). The analysis enables targeted security measures such as customized employee training and improved authentication protocols. For instance, identifying vulnerable departments or roles can prompt additional training (SOCRadar, 2020). Now, even though the calculated budget worldwide for information security has bypassed 114 billion US dollars in 2018 (Steve, 2019), the total number of data breaches has not gone down. Recent breaches have been attempted at Facebook (Winder, 2019), First American Financial Corp (Mathews, 2019), Capital One (Colby, 2019), Uber and many more.

Within the context of these attacks, a predictive model could employ data on previous attacks to identify patterns in user behavior or system vulnerabilities that render an organization more vulnerable to phishing attacks (Orunsolu et al., 2022). However, the issue lies in the inadequate amount of research on the effectiveness of predictive models in detecting social engineering attacks, specifically those involving phishing. While some research has been conducted on the matter, the extent of the knowledge on what attributes to the success of these models remains insufficient as it is centered around general social engineering attacks rather than concentrating on the highly prevalent and hazardous phishing method (Orunsolu et al., 2022).

Consequently, this paper seeks to create and assess a natural language processing (NLP), and machine learning model for detecting phishing attacks facilitated by social engineering. The primary objective is to identify trends and user characteristics associated with these attacks, aiming to enhance the accuracy and efficiency of natural language processing, and machine learning techniques used in combating phishing attacks.

1.2 Research Motivation

Working in Fintech and being able to witness firsthand some of the many cyber-attacks that take place and that are similar across many industries you recognize that something must be missing in the framework organizations use to protect their data. This missing element urged me to perform a literature search to find gaps in the security of financial institutions experiencing the highest amount of cyber threats.

The growing business trend of innovation, however, is threatened by increasing security threats. One major security threat comes from social engineering. It was evident that the industry lacks sufficient database resources to deter cyberattacks which have occurred to create an accurate model, relevant to organizations of different field and size. Social engineering attacks, particularly **phishing**, have become a leading cause of data breaches, resulting in financial losses and reputational damage for individuals and organizations. Per the Q4 2022 report from APWG, out of 1,350,037 total phishing attacks, 27.7% of these attacks were on financial sectors. Social engineering attacks remain a major threat despite increased cybersecurity strategies, underlining the need for innovative and effective approaches to detect and prevent them. These attacks are

multidimensional influenced by actions that come from the intensions and motivations of them (Robinson, et al., 2015).

Many solutions have been proposed to detect phishing attacks. Some of them had been adopted by the industry, which is mainly based on keeping a list of URLs that are expected to be malicious, called a blacklist (ScienceDirect, 2009). However, a main concern about the blacklist approach is the need to continuously update the list by adding newly created phishing websites. And because many of these phishing websites have a short span, and if the process an organization takes to updating the blacklist is slow, then there is a slight chance that a phishing attack can occur without being detected by the blacklist. And phishing through email is one of the simplest forms of cyber-attacks and the most popular.

This paper proposes cutting-edge natural language processing and machine learning model to accurately detect social engineering assaults, notably **phishing** attempts, and identify useful features corresponding to such attacks. This pioneering study may spur the development of powerful remedies to prevent or mitigate social engineering assaults, strengthening information system security.

The findings of this research have the potential to embolden and empower professionals in the field of cybersecurity, IT managers, vulnerable workers, and different stakeholders. The goal is for this model is to help organizations prevent attacks, secure data, and reduce risks. This work may guide future research and inspire innovative social engineering detection and prevention methods. Thus, this research will improve the capacity to recognize and prevent phishing assaults, protecting precious resources and information.

1.3 Problem Statement

Phishing attacks have emerged as a major cyber threat, posing significant risks to individuals, organizations, and their sensitive information. Despite advancements in cybersecurity measures, these attacks continue to exploit human vulnerabilities, resulting in substantial financial losses, reputational damage, and compromised data security. The increasing frequency, sophistication, and effectiveness of phishing attacks demand urgent attention and mitigation strategies to safeguard against this pervasive cyber threat.

An increase in fraudulent activities in financial services and many other business sectors has become an attractive avenue for fraudsters due to innovative expansion to explore new ways to scam them either by gaining access to sensitive information, account takeover fraud (ATO) or New Account Fraud through Synthetic Identity fraud (SIF) (Yeboah-Boateng, & Amanor, 2014).Without proper innovation to protect stakeholders, these attacks will remain an unsolved mystery despite the growth of defense technology in the cyber and technical world. Phishing is yet one of the greatest threats to both personal and business users as these deceptive attempts persist and succeed (Salahdine & Kaabouch, 2019; Albladi & Weir, 2018).

Research to minimize these vulnerabilities across many different sectors could be challenging as each sector could be attacked in a specific way or through a different pathway. The goal is to examine and analyze phishing attacks from different vectors to determine the loopholes and exploit the common weaknesses that personal or business usages make when faced with a phishing attack. As we dive into the ever-evolving world of phishing—an alarming form of cybercrime that exploits social engineering techniques, this paper explores the historical cases, patterns, and characteristics of phishing attacks,

shedding light on their success factors and emerging trends. It emphasizes the adaptability of attackers, who leverage technological advancements to create convincing fake websites and craft persuasive email content and text messages. By analyzing phishing's evolution over time, this research highlights the need for a comprehensive understanding of this cybercrime phenomenon. It uncovers the crucial factors contributing to phishing attacks' effectiveness and demonstrates how attackers exploit psychological triggers like urgency and fear to manipulate victims into disclosing sensitive information.

Presently while numerous clients have the information on what a phishing assault is, they actually fall survivors of these assaults as their insight is restricted to what it is and not how to distinguish it. Therefore, to moderate the consequences associated with phishing attacks, user education is required. It highlights the need to raise awareness among individuals about phishing techniques, deceptive emails, fake websites, and social engineering strategies. Organizations and individuals can develop a vigilant mindset and exercise caution when encountering suspicious online activities by improving user education programs. The use of artificial intelligence (AI) requires additional research targeted on how the phishing attack took place and will help build a more successful defense mechanism to fight this ongoing problem.

1.4 Thesis Statement

The alarming increase of phishing attacks calls for a comprehensive analysis of various forms of such attacks to effectively safeguard sensitive data through the implementation of an NLP model that will combine various phishing attack forms.

There has been a rapid increase in the rate of cyber threats in recent years due to phishing attacks causing a substantial financial loss, sensitive data compromise and blockage of work (Hong, 2012). The goal is to identify trends and user characteristics associated with these attacks in order to improve the accuracy and efficiency of predictive models in combating phishing attacks via any messaging format. This will be accomplished by performing an analysis on various forms of phishing using AI models to determine the quantitative measurements of the datasets and assess a predictive model using NLP for detecting phishing attacks facilitated by different messaging forms.

In summary, the focus of this research will be collecting fraud websites, emails, texts messages, GPT message language, malicious URLs, fraudulent job postings, user information, and several other user data to build a classification algorithm using machine learning. This approach does not need to be updated whenever a new phishing website is introduced. Instead, it can be trained periodically to adjust the classifier for the new phishing attributes.

1.5 Research Objectives

The findings of this research will contribute to developing effective strategies for preventing and detecting phishing attacks. They will have implications for a wide range of organizations and industries. Thus, the objectives are:

- To identify the most common user characteristics associated with social engineering attacks, specifically phishing attacks.
- To evaluate the effectiveness of using different data points in creating a predictive model for detecting social engineering attacks.

- To determine the extent to which data collected from adversaries can be used to identify trends in user characteristics associated with phishing attacks and improve system security.
- To assess the potential of a predictive model as a framework for training employees in industries vulnerable to social engineering attacks.
- To investigate the feasibility of building a predictive model for detecting phishing attacks using a browser extension.

As previously expressed, this paper expects to offer an understanding of the making of productive prescient models that can recognize and forestall social designing assaults, primarily phishing assaults. The examination discoveries will support conceiving methods that work on an association's network protection, shield against social designing assaults, and advance data frameworks' security by and large.

1.6 Research Questions and Hypotheses

This research examines the structure of phishing emails, text messages, fake job postings, and malicious websites. It focuses on identifying specific keywords that exploit psychological triggers and curiosity in potential victims, which are commonly used in phishing attacks. The goal is to develop a more effective structure to prevent such attacks.

R1. Would working with a diverse dataset enhance the ability of the NLP model to detect novel and evolving phishing attacks?

R2. How effective is an NLP-based model at assessing the credibility of email/message/text content to prevent phishing attacks?

R3. How does the performance of an NLP-based phishing detection model compared to traditional rule-based approaches?

H1. An NLP model trained on a vast and diverse dataset can adapt to new phishing tactics and patterns, resulting in improved detection rates and prevention of novel attacks.

H2. By analyzing email/text/message language and employing natural language processing techniques, the model can accurately assess the credibility of the content and thus prevent phishing attacks.

H3. An NLP model, leveraging contextual information and semantic analysis, can outperform rule-based approaches in terms of accuracy and adaptability while detecting and preventing phishing attacks.

1.7 Scope of Research

The world has become a place of constant interest and growth in technology – everyone these days uses a smartphone that contains all their personal information from emails, passwords and banking information. All this sensitive information has a doorway to the online world. What makes phishing so effective stems from human behavior's use of technology to communicate in a fast-paced world. Social engineering attacks, particularly phishing, have become a leading cause of data breaches, resulting in financial losses and reputational damage for individuals and organizations.

Financial cost of data breaches has been reported by Main & Bottorff (2022) to be about 5.3 million dollars in losses while IBMs (2023) report states that the U.S. loss in data breach cost is double that of the world reaching close to 10 million dollars. And 90 % of these data breaches are due to phishing attacks. APWG (2022) fourth quarter statistics show that the observed 1,350,037 over-all attacks are phishing attacks, which is

a novel record and the worst and most critical quarter for phishing attempts that APWG has ever witnessed. Despite being over a decade old, the alarming reality persists that many individuals remain unaware of how phishing operates, making them subject to falling victim to this deceptive scheme.

This proves that even though there is a great deal of research around this topic there is something missing that has adversaries still being able to run these attacks in a successful manner. There needs to be a system in place to help private individuals and organizations obtain a full understanding of these attacks and how to look out for them not only with the help of software but with creating a framework developed from advanced technology using AI and programming languages to detect phishing messages early on.

1.8 Research Limitations

In conducting this research there were several obstacles that were faced. One of the main limitations was the limited availability of phishing data. Phishing attacks are often conducted privately, making it difficult to gain access to real phishing emails and other related data for research purposes or most data centers around malicious URLs. The datasets available for research might not encompass the full spectrum of phishing attacks, making it challenging to develop comprehensive detection techniques that work effectively for all types and variations of phishing attacks.

Phishing techniques continuously evolve to bypass detection measures, making it difficult to stays current with the latest trends and patterns. This limitation can affect the effectiveness of detection methods developed using older datasets. Where in some cases, we might not have access to ground truth labels for the phishing emails they analyze, making it difficult to train and evaluate detection models accurately. Many organizations

and security vendors develop custom phishing detection algorithms but do not release the implementation details or make them accessible to researchers, limiting the ability to compare and replicate their findings.

Some of the other difficulties were in accessing and analyzing the infrastructure used by phishers, such as phishing websites and command-and-control servers, due to legal and technical barriers, limiting the depth of their investigations. Conducting large-scale analysis of phishing attacks often requires significant computational resources, but due to budget constraints or limited access to high-performance computing infrastructure it was difficult to run.

With these limitations in place and because we do not have a universal one format solution to prevent these phishing attacks from happening there is a requirement to implement a more efficient system. With current research focused on the known phishing emails, the goal of this research is basically to attest to a new method using NLP machine learning with the use of multiple data inputs from different vectors that will provide a more accurate rating in detecting phishing attacks in a manual form and conduct further computerization steps in future research.

1.9 Organization of Praxis

The paper covers five chapters: an introduction, literature review, research methodology, results, and discussion and conclusion. The paper's findings have implications for various organizations and industries and offer insights into creating natural language processing (NLP), and machine learning model to detect and prevent phishing attacks.

Chapter 1 introduces the background of the research, the research motivation, the problem, and thesis statement, outlines the research objective and research questions, the hypothesis of the research, and explains the significance of the study and any limitations faced.

Chapter 2 reviews the literature on social engineering attacks, phishing attacks, and predictive modeling in cybersecurity. This section also provides a conceptual framework for the study and highlights the gaps in the existing research.

Chapter 3 of this paper dives into the research methodology applied in this study. It breaks down the intricate details of the data collection process, the variables employed in the analysis, and the statistical approaches adopted to unravel the data.

In *Chapter 4*, the study's findings are laid out. They entail a multifaceted exploration of the user characteristics associated with phishing attacks, including creating a natural language processing (NLP), and machine learning model for detecting them.

Chapter 5 presents a thought-provoking discussion of the study's implications for cybersecurity practice and research. This study's prediction model's possible uses, limitations, and development roadmap are examined here. In addition to the conclusion which summarizes the important results and highlights this study's unprecedented cybersecurity achievements.

— Literature Review

2.1 Introduction

Social engineering attacks are a category of cyberattacks that manipulate individuals psychologically to extract sensitive information or perform actions that jeopardize information system security (Goel et al. 2017). These attacks exploit human vulnerabilities such as trust, curiosity, and fear to obtain unlawful access to systems and the information on them. Social engineering attacks take different forms, including baiting, phishing, pretexting, and tailgating (Chetioui et al., 2022). Among social engineering attacks, phishing attacks are prevalent and pose a significant threat to organizations globally. The personality framework of social engineering is illustrated in Figure 1, along with the characteristics of victims who fall into this trap, to further illustrate the threat posed by this.

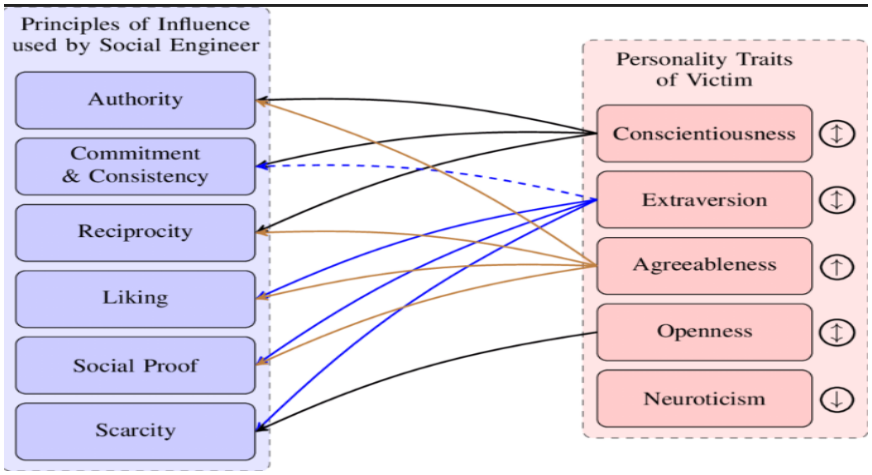


Figure 1: The Social Engineering Personality Framework

Phishing attacks entail using fraudulent emails or messages that appear genuine to deceive recipients into revealing sensitive data or performing actions that compromise

information system security (Goel et al., 2017). The efficacy of these attacks is often due to the ability of attackers to manipulate human vulnerabilities and instill urgency or fear in their targets (Goel et al., 2017). The rising incidence and effectiveness of social engineering attacks are challenging organizations worldwide. Social engineering attacks remain a significant threat despite increased awareness and investment in cybersecurity (Aldawood & Skinner, 2019).

With these attacks' ongoing and unwavering presence, unconventional and state-of-the-art methodologies are paramount to identifying and forestalling such dangers. In this regard, an avenue that appears promising and worthy of exploration is predictive modeling. Harnessing this technique makes it easy to acquire a more thorough perception of the fundamental patterns and trends that characterize social engineering attacks that is conceivable. In other words, predictive modeling creates unique and efficient ways to prevent problems before they happen.

2.2 Review of Social Engineering Attacks

2.2.1 Existing Research on Social Engineering Attacks

Social engineering attacks have plagued organizations and individuals for countless years, leading to a vast amount of research aimed at comprehending the complexity of these attacks and devising ways to prevent them. One area of investigation focused on identifying the most frequent and common types of social engineering attacks. Aun et al. (2023) revealed that the most prevalent are phishing, pretexting, baiting, and quid pro quo attacks. Likewise, Huseynov & Kose (2022) found that phishing attacks are the more common and frequently applied social engineering attack type. This was further examined in a 2020 Verizon study that examined the many causes of data breaches.

Phishing assaults were the most common cause of data breaches, with social engineering coming in second.

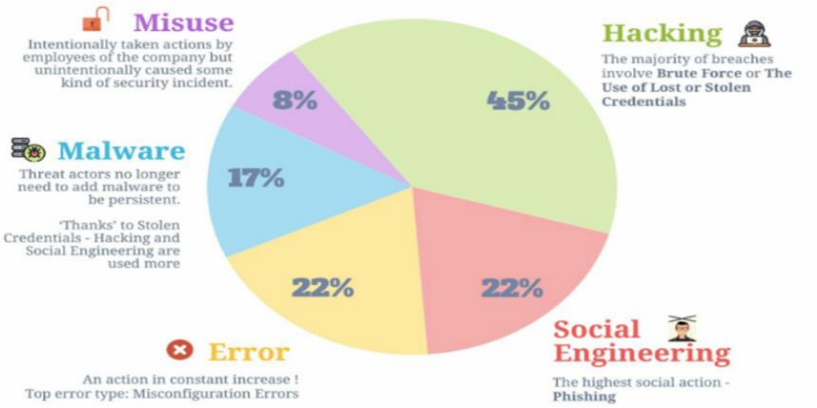


Figure 2: Overview of Verizon 2020 Data Breach Report.

Numerous studies have also concentrated on understanding the distinctive characteristics of individuals most susceptible to falling prey to social engineering attacks. For instance, Montañez et al. (2020) discovered that people who are more trusting, less experienced, and less cautious are likelier to fall victim to such attacks. In the most recent years, there has been expanding interest in developing predictive classification models capable of identifying social engineering attacks before they occur. For instance, Dada et al. (2019) constructed a machine learning-based model for predicting phishing attacks by scrutinizing email header information. Similarly, Tang & Mahmoud (2021) designed a deep learning-based model for detecting phishing attacks by examining user behavior. These models come from understanding the impact of social engineering on the world and this is defined in Figure 3 showing that human error is the number 1 impact to these attacks.

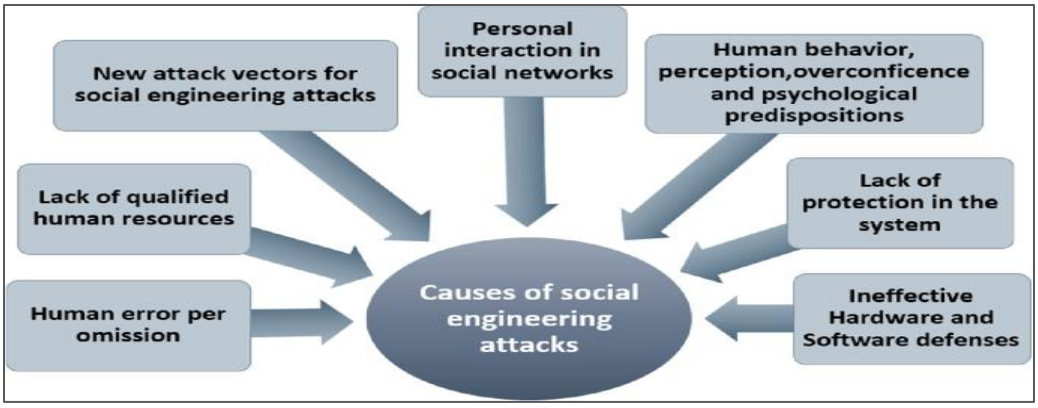


Figure 3: Impact of Social Engineering Attacks

Overall, existing research has provided invaluable insights into the nature of social engineering attacks and their avoidance. Much is yet to be learned, especially as attackers continuously devise new and highly sophisticated methods. Further research is imperative to develop more effective countermeasures and improve the understanding of the factors contributing to the success of social engineering attacks.

2.2.2 **Types Social Engineering Attacks**

Social engineering attacks come in different forms, and attackers utilize various strategies to achieve their aims. Some of the most prevalent social engineering attacks are:

- **Phishing Attacks:** Use fraudulent emails or messages that appear legitimate to deceive recipients into unveiling classified or sensitive information that compromise information systems' security. Phishing attacks employ emotional appeals or persuasive language to sway victims to click on the link or open the file by exploiting their faith in reputable sources (Aura, 2022).

- **Pretexting:** Here the attacker creates a false story or pretext to deceive the victim into revealing sensitive information that compromise information systems' security. The attacker may use fake credentials, phone number spoofing, or create fake websites to establish a convincing pretext (Aura, 2022).
- **Baiting:** It exploits individuals' curiosity and temptation to discover what is on physical media like USB drives or CDs that is found in public places (Aura, 2022).
- **Tailgating:** This involves gaining unauthorized access to a secure area by being physically present through following an authorized person closely (Aura, 2022).
- **Quid Pro Quo:** This attack offers a benefit or reward in exchange for classified or sensitive information to access systems (Aura, 2022).
- **Scareware:** This utilizes fake alerts or warnings to deceive individuals into downloading malware or other dangerous software (Aura, 2022).
- **Watering Hole:** A complex attack that involves compromising a website or online platform that a particular group of individuals frequents to gain access to their sensitive information (Aura, 2022).

2.2.3 How to Avoid or Minimize These Attacks

Understanding the life cycle of social engineering and the four different stages an attacker takes to perform their attack will introduce a better level of exercising caution when examining electronic mail or messages. This is paramount to safeguard oneself from phishing attacks. It is crucial to guarantee the authenticity of the source's electronic

mail address, meticulously inspect the syntax and orthography of the material, and avoid accessing links present in uninvited electronic mail before initiating any form of a hyperlink or revealing classified information (Syafitri et al., 2022). Additionally, ensuring that the software and computerized platforms are routinely updated and consistently utilize anti-malware software to discover and obliterate any pernicious software that may have penetrated one's system is essential.

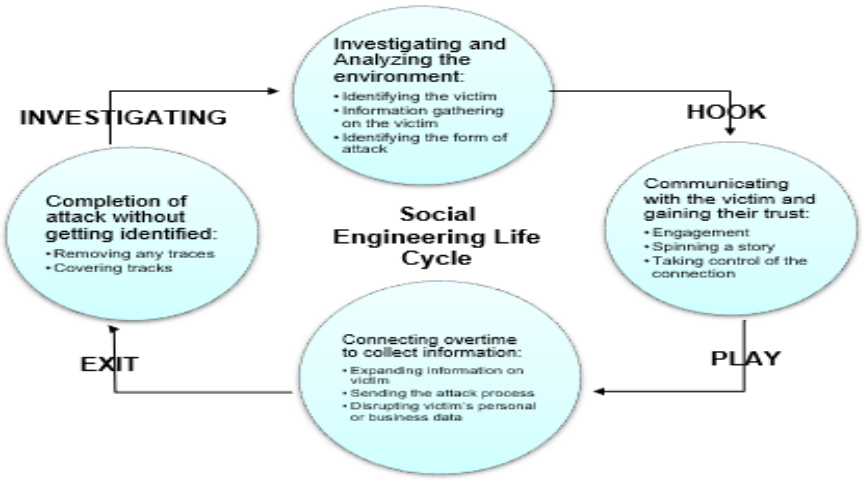


Figure 4: Social Engineering Life Cycle

- ***Pretexting Attacks*** → The best way is to employ numerous communication channels, such as contacting the person or institution using a verified phone number or email address and checking the attacker's credentials and URLs (Syafitri et al., 2022).
- ***Baiting Attacks*** → Be hyper-vigilant of physical media that may be abandoned in public places, especially if it appears suspect or untended. It would help if you also refrained from inserting unknown USB drives or other physical media into your computer while regularly scanning your

system with antivirus software to identify and eradicate any malware (Syafitri et al., 2022).

- ***Tailgating Attacks*** → When entering a protected location, it's important to be conscious of anyone trying to follow you closely. You should also always follow the "need to know" to obtain a valid reason. (Syafitri et al., 2022).
- ***Quid Pro Quo Attacks***→ Familiarize oneself with the policies and procedures of one's organization and to share sensitive information or allow access to information systems only after verifying the requester's identity and legitimacy (Syafitri et al., 2022).
- ***Scareware Attacks*** → One should be vigilant of any pop-up windows or alerts on your computer screen to avoid falling into the trap of scareware attacks. (Syafitri et al., 2022).
- ***Watering Hole Attacks*** → Websites with significant visitor traffic often become targets for hacking, leading to the injection of malware or trojans into these sites. The Water-holing technique patiently awaits the user's interaction, such as clicking on a link or downloading an application from the compromised website, facilitating a two-way communication channel between the hacker and the unsuspecting victim. (Syafitri et al., 2022).

2.2.4 **Techniques Used in Social Engineering Attacks**

Social engineering attacks utilize a myriad of techniques to deceive and exploit individuals' weaknesses. These attacks can be sophisticated and manipulative, using tactics such as authority and scarcity to coerce individuals into revealing sensitive

information that can assist in performing actions that can compromise the security of information systems.

- **Authority Attacks**: Leveraging an attacker's perceived authority or position of power to influence individuals by posing as someone in a position of authority to gain the trust of the target (Hadnagy, 2018).
- **Scarcity Attacks**: Creating a false sense of urgency or scarcity to manipulate the target's behavior or the use of fear tactics, such as threats of financial loss or missed business opportunities, to convince targets to act quickly (Hadnagy, 2018).
- **Familiarity Attacks**: The attacker would instill a false sense of trust and ease by employing trusted language, images, or symbols with which the victim is already acquainted. (Hadnagy, 2018).
- **Social Proof:** The modus operandi behind this attack method is founded on the principle that people are more susceptible to trusting the opinions or actions of others, particularly those who are regarded as similar or influential. (Hadnagy, 2018).
- **Reciprocation Attacks**: Due to the human tendency to reciprocate gestures the attackers may lure their targets with an alluring free trial, a money-saving coupon, or a mouth-watering gift card (Hadnagy, 2018).
- **Fear-Based Attacks:** Fear-based social engineering attacks leverage the power of dread to instill a sense of panic or anxiety in the victim, which can lead to irrational behavior (Hadnagy, 2018).

2.2.5 **Techniques to Thwart Social Engineering Attacks**

- **Authority Attacks →** Verify the legitimacy and identity of any individual who asserts to hold a position of power before acquiescing to their requests. Put into practice regulations and measures that necessitate verification of appeals for sensitive information or actions and provide training to personnel on how to identify and react to authority attacks (Hadnagy, 2018).
- **Scarcity Attacks →** Validate the authenticity of any claims made by the sender and refrain from any actions that could endanger one's information systems' safety without fully scrutinizing the appeal. Institute protocols and measures that demand validation of requests for sensitive information or actions and educate personnel to identify and counteract scarcity attacks (Hadnagy, 2018).
- **Familiarity Attacks →** Verify the authenticity of any appeals made by the sender and avoid engaging in any actions that could jeopardize one's information systems' security without conducting a thorough investigation of the request. Implement regulations and measures that mandate the verification of requests for sensitive information or actions and educate personnel to recognize and counteract familiarity attacks (Hadnagy, 2018).
- **Social Proof Attacks →** Verify the legitimacy of any claims made in the message and avoid engaging in any actions that could jeopardize one's information systems' security without scrutinizing the appeal fully. Institute protocols and measures that require validating appeals for

sensitive information or actions and educate personnel to identify and counteract social proof attacks (Hadnagy, 2018).

- **Reciprocity Attacks →** Validate the offer's and the sender's authenticity before taking action. Put into practice regulations and measures that mandate the validation of appeals for sensitive information or actions and provide training to personnel on how to recognize and react to reciprocity attacks (Hadnagy, 2018).
- **Fear-Based Attacks →** Approach any claims that elicit strong emotional responses with a skeptical mindset and confirm their veracity before taking action. Individuals should take the time to decipher the language used in any messages they receive and seek clarity if necessary. Implement strong passwords, activating two-factor authentication, and ensuring that software and security systems are updated regularly (Hadnagy, 2018).

2.3 Detection of Social Engineering with Predictive Models

There has been an increased degree of enthusiasm for developing predictive models to forecast social engineering attacks before they transpire. These models often employ intricate machine learning or deep learning techniques to scrutinize many data points and unearth patterns that may indicate an imminent attack (Berk et al., 2019). One realm of research has zeroed in on using email header information to predict phishing attacks. For instance, Jupin et al. (2019) devised a machine-learning model that analyzes email header information, such as the sender's email address and IP address, to prophesy phishing

attacks. Astonishingly, the model yielded a staggering 99.45% accuracy in detecting phishing emails.

Other studies have delved into analyzing user behavior to detect social engineering attacks. Hazratifard et al. (2022) introduced a deep learning model that dissects user behavior, such as mouse movements and keystrokes, to uncover phishing attacks. The model delivered a remarkable accuracy of 99.11% in detecting phishing attacks. Apart from these methodologies, several researchers have investigated the use of natural language processing techniques to dissect email content and pinpoint phishing attacks (Alshamrani et al., 2019). For example, a research effort by Salloum et al. (2021) unveiled a model that utilizes natural language processing to scrutinize email content and flag phishing attacks. The model attained a commendable accuracy of 92.16% in detecting phishing emails.

Thus, various research has substantiated the prospective nature of predictive models for detecting social engineering attacks. Nevertheless, several impediments, such as the high-quality data needed to effectively train these models, still linger. Further research is indispensable to perfect these models and bolster their effectiveness in real-world settings.

2.4 Phishing attacks

In today's interconnected world, the advent of computer-mediated communication has undoubtedly revolutionized the way people interact. However, this transformation has come with challenges. One of the most concerning consequences of this digital age is the widespread occurrence of deceptive practices, with phishing as a main and alarming form of cybercrime. Phishing is a clever scheme that capitalizes on social engineering

techniques to manipulate and deceive unsuspecting individuals, ultimately leading them to disclose confidential and sensitive information (Hong, 2012).

Cybercriminals employ various channels by masquerading as trustworthy entities, such as emails, text messages, or voice calls, to lure victims into sharing their data. While phishing has attracted significant attention from researchers, it is continuously evolving nature demands a more comprehensive investigation to keep pace with its intricacies. Ferreira and Marques (2018) assert that a thorough understanding of phishing's evolution over time remains elusive despite extensive research efforts. Exploring historical cases and analyzing their patterns and characteristics is crucial to gain deeper insights into this cybercrime phenomenon. By examining past instances, researchers can identify essential factors contributing to phishing attacks' success and shed light on emerging trends and techniques (Ferreira and Marques, 2018).

Schuetzler et al. (2011) highlights the evolution and diversification of phishing attacks, which have adeptly capitalized on technological advancements and vulnerabilities inherent in computer-mediated communication channels. Attackers have adopted various strategies to deceive their targets, leveraging tactics like creating fake websites that closely mimic legitimate platforms. Additionally, they craft persuasive email content and employ enticing links in text messages (Schuetzler et al., 2011). Moreover, these cybercriminals adeptly manipulate psychological triggers, exploiting factors such as urgency or fear to coerce individuals into disclosing their personal information.

As computer-mediated communication continues to advance, so do the tactics employed by phishing attackers. Schuetzler et al. (2011) emphasizes the adaptive nature

of these cybercrimes, which exploit the ever-expanding array of technological tools. Attackers have honed their skills in creating deceptive websites, meticulously designing them to resemble authentic platforms. These counterfeit sites can easily mislead unsuspecting victims into sharing sensitive information, as they are virtually indistinguishable from legitimate websites. In addition to deceptive websites, phishing attackers have perfected the art of crafting compelling email content. They skillfully manipulate language and employ psychological techniques to create a sense of trust and urgency. By using persuasive language and appealing to emotions, attackers exploit individuals into revealing their personal information willingly. Text messages have also become a preferred channel for phishing attacks, as they offer a direct and immediate line of communication with potential victims (Alkhalil et al., 2021).

Attackers utilize enticing links in text messages, often designed to provoke curiosity or offer enticing rewards. These links, when clicked, can lead to malicious websites, or initiate the download of harmful software onto the victims' devices. To maximize their chances of success, phishing attackers exploit psychological triggers. By leveraging factors such as urgency or fear, they create a sense of pressure that compels individuals to act without careful consideration. For example, attackers may send urgent emails claiming account suspensions or security breaches, instilling panic, and prompting victims to disclose their personal information hastily (Alkhalil et al., 2021). The phishing process is broken down in Figure 5 showing how the adversary focuses on a target, prepares the right attack method and goes into the attack to obtain sensitive information.

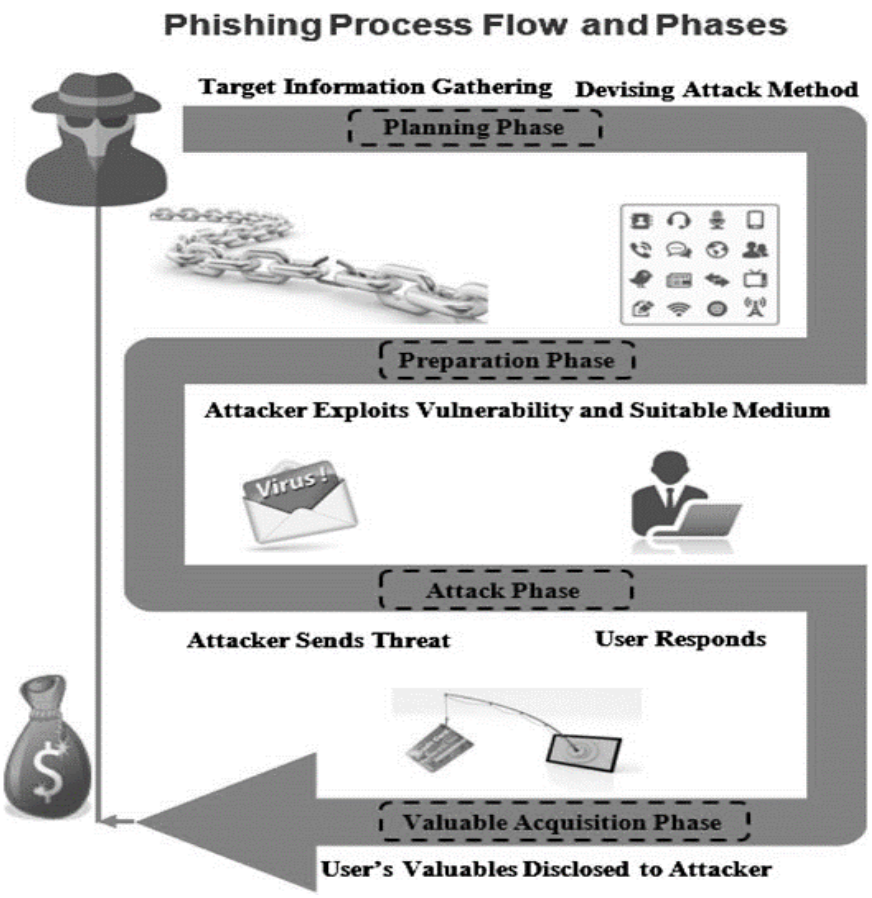


Figure 5: Phishing Process Flow and Phases

The effectiveness and widespread occurrence of phishing attacks have been significantly facilitated by the accessibility of technical resources readily available from both public and private sources (Milletary, 2005). The availability of these resources has contributed to lowering the entry barrier for individuals with limited technical expertise, enabling them to partake in these illicit activities. By automating certain aspects of phishing campaigns, attackers have simplified the process and made it more accessible to a broader range of individuals (Milletary, 2005). Automating specific tasks within phishing campaigns has played a crucial role in expanding the reach and impact of these attacks.

Previously, phishing required specific technical proficiency, making it a specialized field. However, with the advancement of technology and the availability of resources, attackers have been able to automate various stages of their campaigns. This automation includes tasks such as the creation of phishing websites, the distribution of phishing emails or text messages, and the collection of stolen data. By automating these processes, attackers can conduct phishing attacks at scale, targeting a more significant number of individuals simultaneously (Milletary, 2005).

Moreover, attackers continuously adapt their tactics to exploit new technological developments and circumvent existing security measures (Milletary, 2005). Cybercriminals innovate and find new ways to bypass these defenses as security measures and technologies evolve to combat phishing attacks. They remain vigilant in monitoring advancements in security protocols, software vulnerabilities, and user behaviors to identify weaknesses that can be exploited. This adaptive nature enables them to stay one step ahead of security measures, making it a constant challenge for organizations and individuals to protect themselves effectively (Milletary, 2005).

The proliferation of technical resources and the automation of phishing campaigns have transformed phishing into a widespread and persistent threat. The accessibility of these resources has expanded the pool of potential attackers, allowing less technically proficient individuals to engage in this illicit activity.

Furthermore, the adaptability of attackers to exploit new technological developments showcases the ever-evolving nature of phishing attacks. To effectively combat this threat, researchers, organizations, and security professionals need to remain proactive in their efforts, continually enhancing security measures and educating users about the risks

associated with phishing. User behavior and awareness play a crucial role in the success of phishing attacks Yeboah-Boateng & Amanor (2014). They found that individuals with limited knowledge of phishing techniques and low awareness levels are more susceptible to falling victim to these attacks. Abraham & Chengalur-Smith (2010) also observed that gender differences have also been observed, with men exhibiting higher levels of trust and susceptibility than women. Improving user education and awareness programs is essential to mitigating the risks posed by phishing attacks.

Thus, minimizing the detrimental impact of phishing attacks necessitates a comprehensive approach that integrates technological solutions and user education (Alseadoon et al., 2015). Technological countermeasures are pivotal in combating phishing attempts by employing anti-phishing tools, spam filters, and browser extensions (Gupta & Kumaraguru, 2014). These tools are designed to identify and block phishing attacks, providing a crucial line of defense. However, given the constantly evolving tactics employed by attackers, it is imperative to continually update and enhance these technological countermeasures (Ferreira & Marques, 2018). Developing robust anti-phishing tools is vital to effectively detecting and thwarting phishing attempts. These tools employ advanced algorithms and heuristics to analyze incoming emails, messages, or website links. By comparing them against known phishing patterns and indicators, these tools can identify suspicious content and alert users or automatically block access to potentially harmful websites.

On the other hand, spam filters play a significant role in identifying and diverting phishing emails to users' spam folders, minimizing their chances of being deceived. Furthermore, browser extensions can provide an added layer of protection against

phishing attacks. These extensions integrate with users' web browsers, offering real-time analysis and alerts for potentially malicious websites. They can detect phishing indicators, such as forged URLs or suspicious pop-ups, and warn users before interacting with potentially harmful content. While technological countermeasures are crucial, user education programs are equally important in mitigating phishing attacks (Alseadoon et al., 2015). These programs aim to raise awareness among individuals about the tactics employed by attackers and provide them with the necessary knowledge and skills to identify and respond effectively to phishing attempts (Ferreira & Marques, 2018). Individuals can develop a vigilant mindset and exercise caution when encountering suspicious online activities by educating users about standard phishing techniques, such as deceptive emails, fake websites, or social engineering strategies. Continuous evaluation and refinement of technological solutions and user education programs are imperative to keeping pace with the evolving nature of phishing techniques.

Regular assessments of the effectiveness of anti-phishing tools, spam filters, and browser extensions allow for necessary updates and improvements (Alseadoon et al., 2015). Similarly, user education programs should be evaluated to ensure their relevance and effectiveness in addressing emerging phishing trends and tactics. Organizations and individuals labeled below in Figure 6 show the organizations by the most targeted sectors attacked via phishing. They can enhance their resilience against phishing attacks and reduce their potential impact by adopting a multifaceted approach that combines technological advancements and informed users.

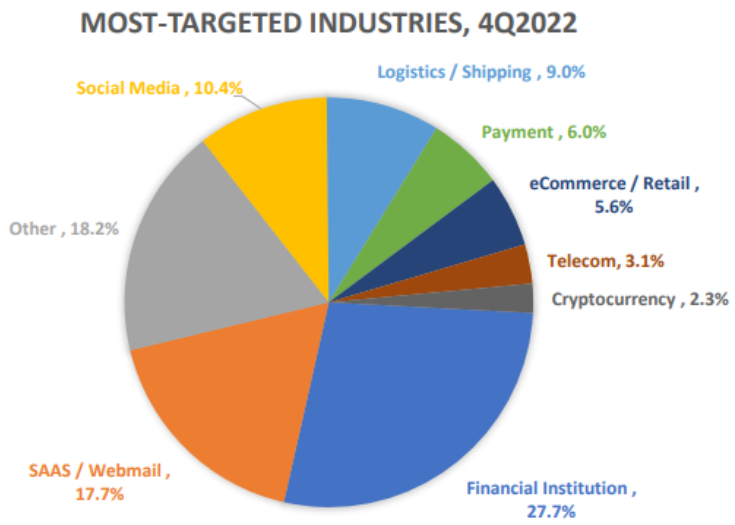


Figure 6: Most Targeted Phishing Sectors

2.4.1 Types of Phishing Attacks

There are several Phishing attacks where each is designed to deceive individuals into revealing sensitive information:

- **Deceptive Phishing** - Emails or messages are sent impersonating a legitimate organization or person, aiming to trick recipients into providing personal information, such as passwords, social security numbers or even financial information like credit card details (Bhavsar et al., 2018).
- **Spear Phishing** - They gather personal information about their targets to craft personalized and convincing emails, making it more likely for victims to believe the message is legitimate (Bhavsar et al., 2018).
- **Whaling** - They target high-profile individuals such as executives or high-ranking employees (Bhavsar et al., 2018).
- **Vishing** - Short for "voice phishing," involves attackers using phone calls to deceive victims (Bhavsar et al., 2018).

- **Smishing** - Smishing attacks employ SMS (Short Message Service) or text messages to trick victims into clicking a malicious link or providing personal information. The messages appear to be from a trusted source inducing victims to take immediate action (Bhavsar et al., 2018).
- **Pharming** - This involves redirecting victims to a fake website that resembles a legitimate one. Attackers manipulate DNS (Domain Name System) or modify hosts files to redirect users to fraudulent websites (Bhavsar et al., 2018).
- **Man-in-the-Middle (MitM) phishing:** The attackers intercept and alter the communication between two parties, making both believe they are communicating directly with each other. They can steal information by eavesdropping or modifying the data transmitted, leading victims to disclose sensitive details (Bhavsar et al., 2018).
- **Clone phishing:** The attackers create a nearly identical copy of a legitimate email or website and replace a legitimate attachment or link with a malicious one (Bhavsar et al., 2018).

2.4.2 Detecting Phishing Attacks

There are several techniques that are commonly used to detect phishing attacks:

1. **Email filtering and analysis**: Many email service providers use advanced filtering algorithms to identify and block phishing emails. These algorithms analyze multiple aspects including sender reputation, email content, attachments, and embedded URLs to flag suspicious emails (Bhavsar et al., 2018).

2. **URL analysis**: Various tools and services can analyze URLs found in emails or web pages to determine if they are legitimate or potentially malicious. These tools check for known phishing indicators such as misspelled domains, deceptive URLs, or redirects to suspicious websites (Bhavsar et al., 2018).
3. **SPF, DKIM, and DMARC:** These are email authentication protocols used to authenticate the legitimacy of the sender's domain. Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM), and Domain-based Message Authentication, Reporting, and Conformance (DMARC) help detect forged or unauthorized emails by checking the email's source against established policies (Bhavsar et al., 2018).
4. **User awareness and training**: Educating users about common phishing techniques and providing regular awareness training can help in detecting and preventing phishing attacks (Bhavsar et al., 2018).
5. **Phishing intelligence and blacklists**: Security companies and organizations maintain databases of known phishing websites and email senders. These lists are regularly updated and used by various security tools to compare and detect potential phishing attacks (Bhavsar et al., 2018).
6. **Pattern recognition and anomaly detection**: Advanced algorithms and machine learning techniques can be employed to analyze patterns and detect anomalies in email content, headers, or user behavior. Any deviation from typical communication patterns can indicate the presence of a phishing attack (Bhavsar et al., 2018).

7. **Real-time monitoring and reporting**: Real time continuous monitoring of network traffic, email communications, and user activity can help identify and respond to phishing attacks promptly (Bhavsar et al., 2018).

Description	Pros	Cons
Detects phishing attacks by using a whitelist filter.	* Pages that bypass the whitelist filter are filtered again by Support Vector Machines. * Maintains accuracy of whitelist filter by using a personalized whitelist.	* Limited dataset of 850 pages. * Unable to detect the attachment of DNS spoofs to legitimate web pages. * High False positive rate.
Implement a comment spam detection mechanism that can be used as a browser plugin and remove spam comments.	* Balances dataset by applying WEKA filters to get the best suitable features. * Spam detection classifier can accommodate new features and detect new classes of spam content.	* Does not do well with a random dataset without applying a supervised resample filter.
Proposes a machine learning-based method that can detect whether a web page exhibits phishing attacks.	* Proposed method is based on an easy to acquire feature vector that does not require additional computation.	* Only uses 10 features for detection. * Limited dataset of 1353 instances.
Uses feature selection to identify important features that categorize phishing and legitimate websites.	* Feature selection highly improves the accuracy score after implementation. * Use of feature selection reduces computational time.	* 14 features. * limited dataset (200 legitimate URL and 1400 phishing URL) * May not work properly with datasets of equal URLs of legitimate and phishing web pages.
Builds a system using machine learning that can classify websites using URLs.	* Can be used to build a rule-based system with associative rules to classify URLs.	* 9 features for each URL * All features are discrete. * Limited dataset (1353 URLs)
Proposes a learning-based aggregation analysis mechanism to decide page layout similarity, which is used to detect phishing pages.	* Automatically trains classifiers to determine web page similarity from CSS layout features, which does not require human expertise.	* Method is lightweight as it only takes one class of features, CSS structure. * Limited by the size of the dataset and distribution of samples.
This research uses a new attribute called the "domain top page similarity" to improve the efficiency of a machine learning-based phishing detection model.	* Increases f-measure and reduces the error rate. * Proves that with better features, the detection rate is much higher and can be implemented in future works.	* The model is highly dependent on the accuracy of the features.
This paper proposes a real-time anti-phishing system that uses seven classification algorithms and natural language processing-based features (NLP)	* Independence from language and third party services. * Huge dataset of legitimate and phishing data. *Real-time execution. * Can detect new websites because of NLP features.	* Machine learning-based systems cannot correctly utilize such a vast dataset.
Performs an extensive measurement of squatting phishing, where the phishing pages impersonate target brands at both the domain and content level.	* Uses features from visual analysis and optical character recognition. * Open sourced tool. * Uses evasive behaviors of phishing pages to build classifiers.	* Unable to detect phishing pages that use cloaking. * Only focuses on popular brands. * The classifier cannot be compared with other phishing tools like CANTINA and CANTINA+.
Uses features from HTML content, JavaScript code and URLs to build a classifier that can detect malicious web pages and threat types.	* Diverse features. * High accuracy score. * Highlights features that are necessary to extract.	* Limited dataset (2500 URLs) * Classifier may not do well with large datasets.

Table 1: Comparison of Machine Learning Based on Phishing Detection Systems

It is important to note that while these techniques aid in detecting and mitigating phishing attacks, no method is foolproof. Combining multiple approaches and promoting user vigilance is crucial in staying safe from phishing threats (Alam et al., 2020).

2.4.3 Anatomy of Phishing Messages

Phishing messages are deceptive messages created with the intention of tricking recipients into revealing sensitive information that create a pathway for actions to be

performed that can compromise their own and system's security. Think of the messages as a prey that survives on human emotions and behavior. It has been surveyed that most of the phishing attacks are successful as they are messages that the individual is waiting for or anticipating receiving Wombat Security (Egan, 2016). Siddiqi et al. (2022) discuss that most adversaries imitate real names of individuals or organizations that the victim is aware of, as well as similar domains and URLs, to deceive the victim into interacting with phishing emails and clicking on the given URL.

In Figure 7, there are nine samples of phishing emails. These typically consist of several components carefully crafted to manipulate and persuade individuals. Here are some key components and examples of phishing messages:

1. **Sender Identity**: Phishing messages often forge the sender's identity to appear as a well-known and trusted entity. These can include imitating reputable companies, financial institutions, or government organizations. For example:
    - An email claiming to be from "PayPal Security Team" asking the recipient to update their account details urgently.
2. **Urgency and Fear**: Phishing messages often create a perception of urgency or fear to provoke immediate action without proper consideration. Examples include:
    - A message declaring that the recipient's bank account will be closed unless they log in and verify personal information.
    - An email claiming unauthorized access to an online account, instructing the recipient to change their password immediately.

3. **URL Spoofing**: Phishing messages frequently contain disguised or misleading URLs to direct recipients to fraudulent websites that mimic legitimate ones. For instance:
    - An email claiming to be from a popular online retailer, with a masked URL that redirects to a counterfeit website designed to steal login credentials.
4. **Social Engineering:** Phishing attacks often leverage psychological manipulation techniques to persuade individuals to take action. Examples include:
    - A message impersonating a colleague or superior, requesting the recipient to disclose sensitive corporate information.
    - An email proclaiming the recipient as the "lucky winner" of a lottery and demanding personal information to claim the prize.
5. **Grammatical and Spelling Mistakes**: While not always present, phishing messages often contain grammatical errors or misspellings, which can indicate suspicious activities. For example:
    - An email claiming to be from a reputable institution with multiple grammatical errors and strange sentence structure.
6. **Request for Personal Information**: Phishing messages typically aim to acquire personal information such as credit card details, passwords, social security numbers, etc. Examples include:
    - A message asking the recipient to provide their username and password for an alleged system update.

- An email requesting credit card information under the pretext of verifying billing details.

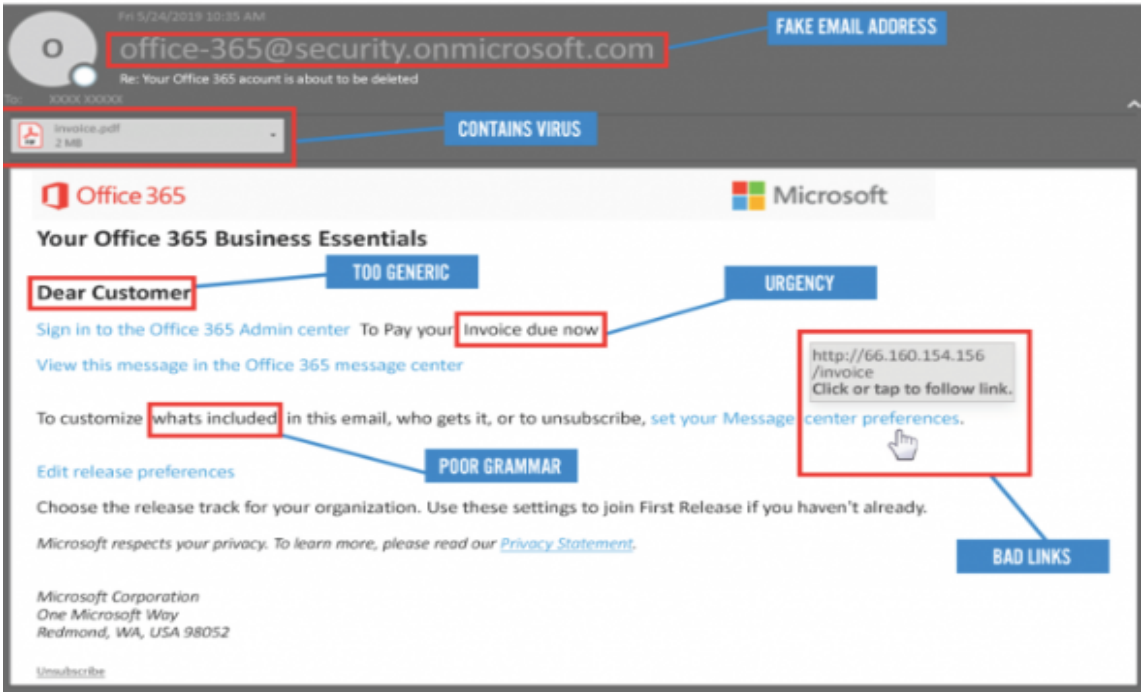


Figure 7: Phishing Email Anatomy for Microsoft Office

In Figure 7, the sender's email address does not appear to be a Microsoft email address. The sender greets the receiver in a very generic way and types a message with poor grammar. The urgency of the email using the key word "NOW" makes this email a suspicious one (Bryley, n.d.).

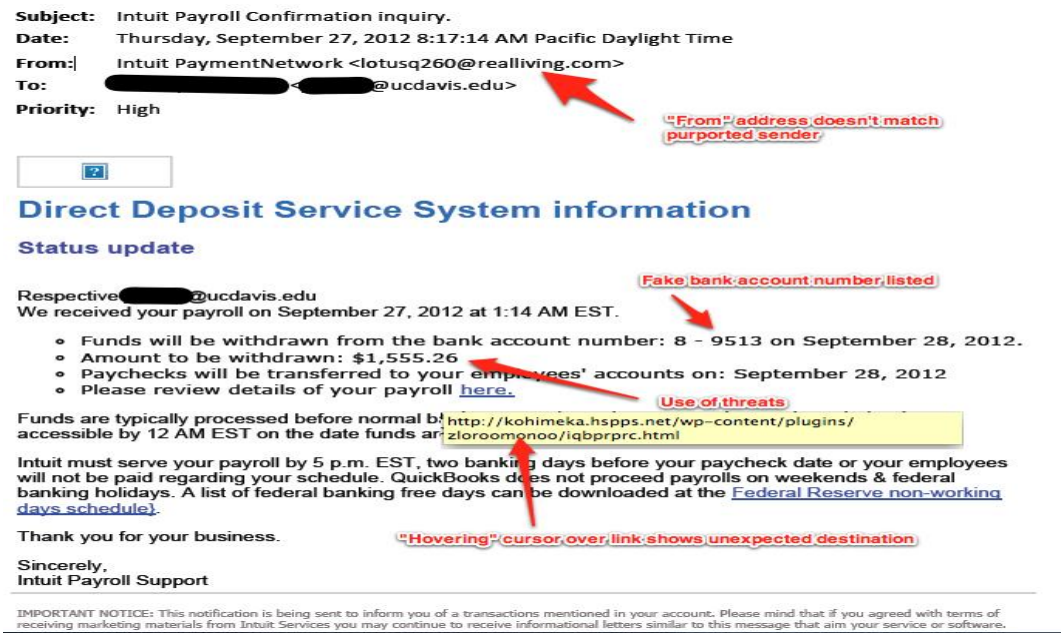


Figure 8: Phishing Email Example for Payroll Setting

Figure 8 is a sample of an email from payroll requesting information to proceed with your direct deposit. This email should be from HR if valid and you can see in the email address that it is not provided from HR raising a red flag (Bryley, n.d.).

Figure 9: Phishing Email Example from the IRS

Figure 9 displays an email from the IRS – this email shows a sense of authority which makes the victim. However, when hovering over the link it is obvious that it is not from the IRS and if clicked will be taken to a malicious website that may hack your system. compile to it as it seems to be an urgent email with the IRS information and logo (Bryley, n.d.).

Cybercriminals integrate the social influence of fondness into their attacks through phishing emails, websites, recruitment ads, malicious URLs and text or voice messages, by attracting their targets through the target's interests and beliefs. It is crucial to remain vigilant and cautious when encountering such messages, scrutinizing their contents, and verifying their authenticity before taking any action.

2.4.4 **Impact and Statistics of Phishing Attacks**

Phishing attacks are only increasing and becoming more sophisticated. Per Violino (2023) there was a 61% increase in the total rate of phishing attacks in the last six months of 2022 ending with Q3 compared to the 2021 that was analyzed by SlashNext. They analyzed billions of messages, attachments, emails, URLs that showed over 255 million attacks took place in 2022. This proves that the attacks are no longer coming from email messages but through text messages, fake posts and even voice calls (Fedor, 2023).

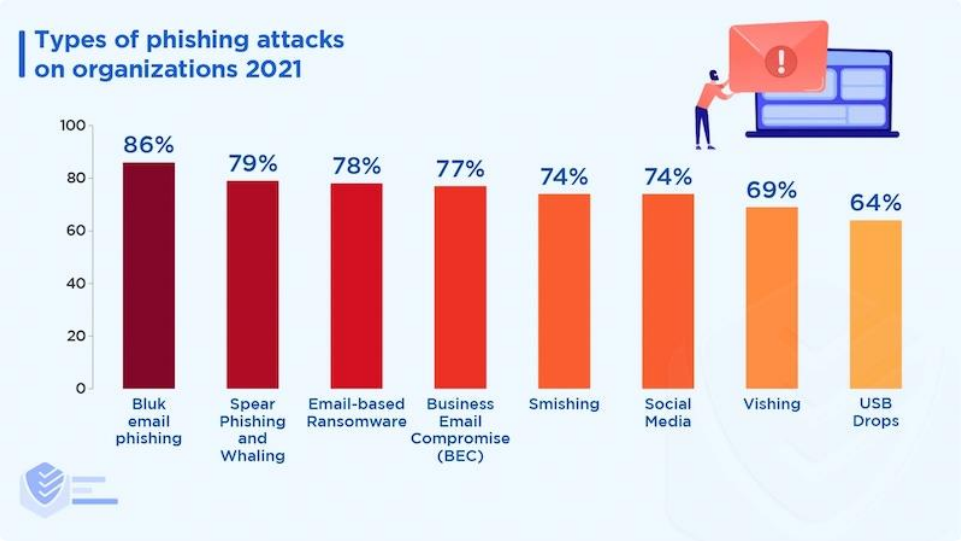


Figure 10: 2021 The Sources of Phishing Attacks

The minds of cybercriminals keep evolving and their attack tactics change with every anti phishing procedure placed in an organization or even on a personal level. According to Q4 of APWG (2023) reports the phishing attacks hit the highest ever seen where there were more than 4.7 million attacks in just 2022 where a third where from Q4. This shows that these attacks are only growing with time. In October alone about 100k phishing emails were detected where each one was different than the other. This is proof of how the criminal minds of these adversaries are evolving.

Additional stats were done by other cyber companies, and it was analyzed by Cofense (2023) in Q1 of 2023 that loaders were the most common form of phishing followed by keyloggers and stealing of information showing that stealing credentials is no longer the number one response that adversaries are looking for when performing a phishing attack.

The change in tactics when performing phishing attacks has made it difficult to stop the game of phishing as the emails, texts or any form of phishing communication is tailored to the victim making it look more realistic. Many fake logins are hosted on Microsoft Azure or through the BEC – where business emails are compromised.

According to APWG Q4 2022 reports it shows tha an average BEC attack that took place attempted to steal $132,559 where 31% came from fraudulent gift card requests. Sharepoint, AWS, and even Google are the most common login domains that have been used to perform phishing attacks (Cofense 2023). Now with the growth of more social media platforms, attackers are using these platforms like YouTube, TikTok, Instagram etc. to add a fake link for followers to click on and become their bait.

With the growth of AI attackers are spam calling individuals recording their voices and using AI to replicate their voices to use as a bait to capture the victims. Zacharakos (2023) reports that ChatGPT for example requires little coding knowledge to help create fake accounts and create malicious code to run a cyber-attack. But since AI is man-made, I can reverse program the tools to detect phishing attacks whether they are coming from texts, messages, and voice communications. Securelist by Kaspersky in 2023 found that ChatGPT can detect phishing URLs about 64% of the time so imagine if programmed correctly how much percentage it can be detected (Tushkanov, 2023).

Financial institutions are still one of the highest targeted industries, but attackers now use tactics to obtain the financial information through other platforms like streaming services, online games, GitHub accounts, etc and doing this using mobile platforms. People have everything stored on their smartphones and there has been an increase of over 50% of credential theft (Fedor, 2023).

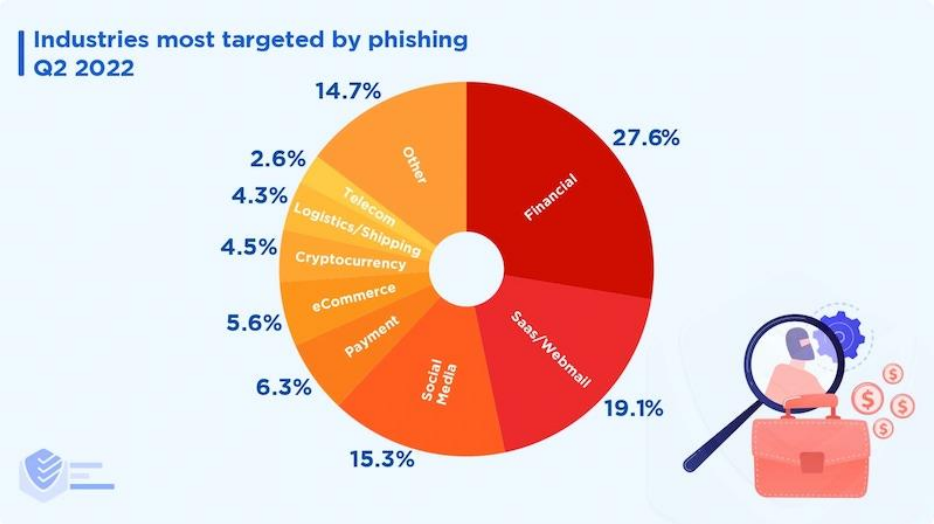


Figure 11: Industries Targeted by Phishing Attacks

2.5 Machine Learning-Based Methods

2.5.1 Current Phishing Detection Models

To counter phishing threats, researchers have explored machine learning techniques as an effective approach for detecting phishing attacks. Several studies have highlighted the effectiveness of machine learning algorithms in distinguishing legitimate websites from phishing attempts. For instance, Dutta (2021) compared multiple classification algorithms and demonstrated that Support Vector Machines (SVMs) achieved the highest accuracy in detecting phishing websites. Similarly, Liu & Chen (2019) utilized a hybrid model combining Random Forest and Principal Component Analysis, achieving a superior detection rate compared to individual algorithms. These studies suggest that machine learning can effectively contribute to the detection of phishing attacks by analyzing various features and patterns associated with phishing behaviors (Basit et al.,2021).

The machine learning methods employed in the most widely used feature detection systems, such as Artificial Neural Networks (ANN), Graph Convolutional Networks

(GCN), Bayesian Additive Regression Trees (BART), and Natural Language Processing (NLP), are the subject of other well-known and recent research on phishing detection. (Mittal et al., 2022). Early research by Ramanathan & Wechsler, (2012) proposed a methodology of multi-layers that was called the phishGILLNET. Where in this research the application of the methods was performed three times through: Fisher similarity, then Adaboost, and finally used the NLP techniques on words that are misspelled to help expand on identifying phishing. This was later expanded by Sahingoz et al. (2019) where they added an additional layer that included classification algorithms and natural language processing base features. This was done to be able to identify malicious links and emails in the present time instead than obtaining it from a list of databases. But even with this there was still a missing component as the models were based on one approach.

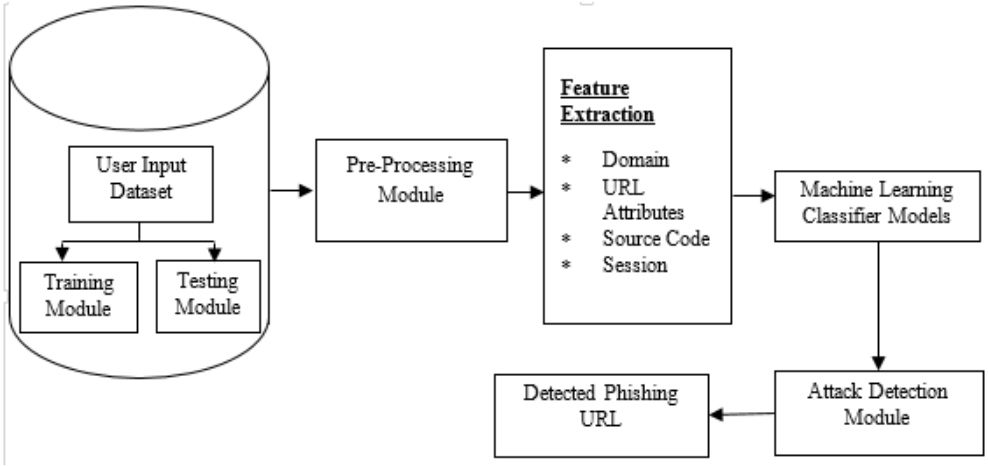


Figure 12: One Approach Machine Learning for Phishing Attack Detection

Several other machine learning models have been used to detect phishing attacks, but none of them can fully prevent them from happening. Some of the previous models used for phishing detection include:

- ***Rule-based models***: These models were based on manually defined rules and patterns for identifying phishing emails. They typically used

features such as URLs, email headers, and suspicious keywords. However, rule-based models are limited in their ability to adapt to evolving phishing techniques, as new patterns and tactics can easily bypass them.

- ***Heuristic models***: These models relied on a set of heuristics or predefined rules to classify emails as phishing or legitimate. They considered features like email sender, content, and payload characteristics. While heuristic models could detect basic phishing attacks, they often had a high false-positive rate and struggled to identify sophisticated attacks.
- ***Machine learning classifiers***: Various classifiers such as decision trees, random forests, logistic regression and support vector machines have been used to detect phishing attacks. These models were trained on labeled datasets containing phishing and legitimate examples, learning patterns and features to differentiate between the two. However, phishing attacks constantly evolve, and ML models often struggled to generalize well to unseen attack variations.

The limitations of previous machine learning models in preventing phishing attacks can be attributed to several reasons like the continuous evolving tactics. Phishing techniques continuously evolve, with attackers employing new strategies, social engineering techniques, and obfuscation methods. Machine learning models trained on former data might resist to detect new attack patterns that were not present in their training set.

Attackers actively try to bypass machine learning models by manipulating input data or leveraging adversarial techniques. For instance, they might obfuscate malicious URLs or attempt to manipulate features used by the models, making it difficult for the classifiers to generalize accurately.

Another limitation can be due to the imbalanced datasets when working to analyze the case. Constructing and maintaining labeled datasets for training phishing detection models can be challenging. Phishing attacks are relatively rare compared to legitimate emails, leading to imbalanced datasets. ML models trained on imbalanced data can result in biased classifiers and decreased performance. There is also a lack of dynamic updates. Previous models often lacked the ability to update themselves dynamically. Phishing attacks occur in real-time, so models trained on outdated data might not be effective in the face of novel attacks.

However, while machine learning-based techniques show promising results, some researchers also raise concerns regarding their limitations. For instance, Tahir et al. (2016) emphasized the need for continuous updating of machine learning models as attackers evolve their techniques. In addition, the reliance on large-scale datasets and the potential for false positives are challenges that require further investigation. Addressing these limitations requires continuous research and development of more robust and adaptive machine learning models and integration of multiple detection techniques, including anomaly detection and behavioral analysis (Tahir et al., 2016).

2.5.2 **Phishing Detection Framework**

Phishing detection frameworks can vary from one organization to the another and can significantly mitigate risks, therefore, it is crucial to remain vigilant, continuously

educate users, and stay updated on the latest phishing tactics whether at home or business. Most frameworks contain 8 analyses, described below, that will help provide a stronger security system against phishing attacks.

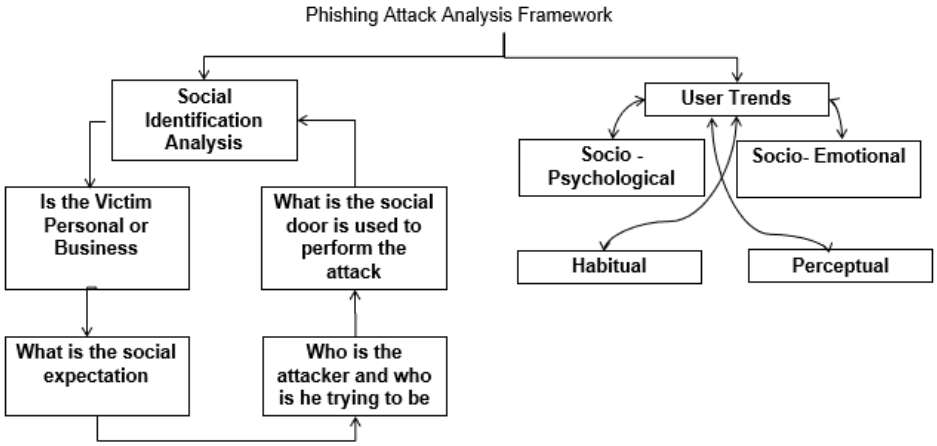


Figure 13: The Analysis Framework for Phishing Attacks

1. Email Header Analysis: Verify the sender's email address and ensure it matches the domain of the organization. Check for suspicious email headers like misspellings, unusual characters, or variations in domain names (Lötter & Futcher, 2015).
2. URL Analysis: Look for URL inconsistencies or misspellings that may redirect users to a malicious site. Hover over hyperlinks to preview the actual URL destination before clicking (Lötter & Futcher, 2015).
3. Content Analysis: Look for poor grammar, spelling mistakes, or excessive use of urgency to create a sense of urgency and panic. Be cautious of email attachments, particularly from unknown sources (Lötter & Futcher, 2015).
4. Sender Analysis: Verify the legitimacy of the sender's email address by cross-referencing with known contacts or official sources. Beware of emails

claiming to be from a reputable organization but using generic email addresses like Gmail or Yahoo (Lötter & Futcher, 2015).

5. Phishing Reporting: Educate users on how to report phishing emails to the appropriate IT department or security team. Implement an easy-to-use reporting mechanism within email clients to report suspicious emails (Lötter & Futcher, 2015).
6. Employee Training: Employees should receive regular security awareness training as a means of teaching them about phishing tactics and warning signs to look out for. Provide guidance on best practices for handling suspicious emails and what actions to take if they suspect phishing (Garera et al., 2007).
7. Anti-Phishing Software: Implement anti-phishing technology that can automatically detect, and block known phishing emails. Use email filters and spam detection mechanisms to reduce the likelihood of phishing emails reaching the inbox (Garera et al., 2007).
8. Ongoing Security Monitoring: Continuously monitor network traffic and email communications for signs of phishing attacks. Stay informed about emerging phishing techniques and adapt the detection framework accordingly (Garera et al., 2007).

2.5.3 **Human and Organizational Error**

In phishing attacks, human and organizational errors refer to mistakes or failures made by individuals or entities that contribute to the success of the attack. The human errors can occur through clicking on phishing emails or links as humans may unknowingly click on malicious emails or links that lead them to fake websites, allowing

attackers to steal their credentials or personal information. Individuals may share sensitive information such as passwords, account details, or personal data with unauthorized individuals through phishing scams (Desolda et al., 2021).

Attackers often manipulate individuals through social engineering techniques to convince them to disclose confidential information or perform certain actions. And if they get enough information the attacker can try to hack an account if the victim uses weak passwords or reuses them across multiple accounts, making it easier for attackers to gain unauthorized access.

There must be a continuous security awareness process for employees and individuals who are not adequately aware of phishing attacks and their consequences as they are more likely to fall victim to such attacks (Desolda et al., 2021). This is done when organizations have specialized teams to build, monitor and educate on security measures. Robust security measures are required as a lack of advanced email filtering systems or effective spam detection makes it easier for phishing emails to reach employees. And with a poor cybersecurity training and education for employees it will leave them vulnerable to phishing attacks.

Organizations that do not have clear reporting channels or communication mechanisms in place may delay the identification and response to phishing attempts. And without a strict data protection protocol many may fail to handle sensitive information with care making it easier for attackers to carry out successful phishing attacks. That is why software, systems and applications must all be updated with the latest security patches as negligence of it can leave organizations susceptible to phishing attacks that exploit known vulnerabilities.

2.6 User Behavior

Phishing attacks have grown in sophistication, making it challenging for traditional detection methods to keep up. Consequently, researchers have turned to behavioral approaches as a means of tackling this problem. Several studies have explored the use of user behavior analysis to detect phishing attacks. For instance, Alkhalil et al., (2021) conducted a comprehensive investigation of mouse movement patterns and demonstrated that certain characteristics can differentiate between legitimate users and those under attack. Similarly, Venkatesha et al. (2021) examined keystroke dynamics and found that unique typing patterns could help identify phishing attempts. While the work of Dong et al. (2008), as seen in Figure 14, worked on a study attempting to reveal phishing websites/URLs. This process was constructed from the sites that the users commonly visited and any information they had gained or entered on these websites. Kumar et al. (2020) introduced an innovative method for phishing email detection, combining SVM with a probabilistic neural network to create a hybrid approach. These studies indicate that analyzing user behavior offers potential for early detection and prevention of phishing attacks.

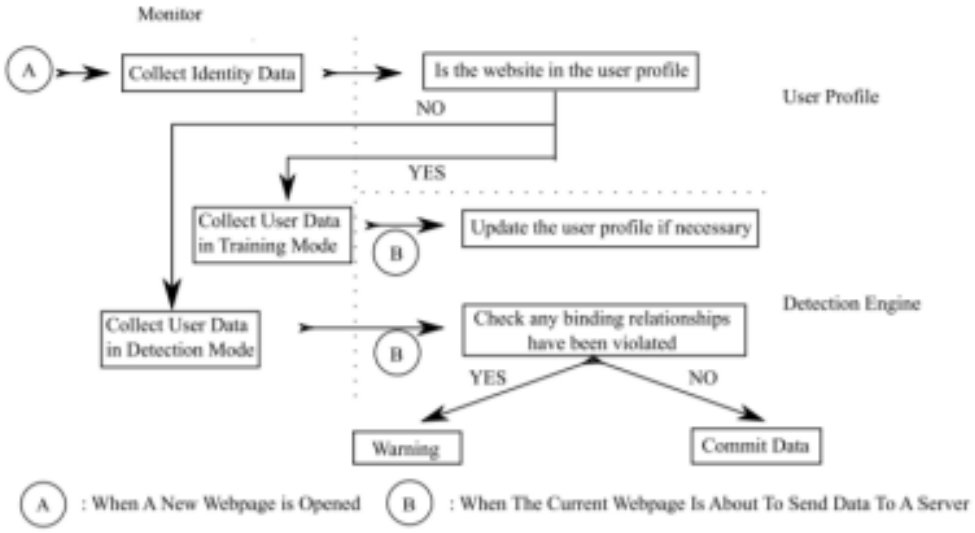


Figure 14: Detecting Website

Source: (Dong et al., 2008)

Adversary behavior in phishing attacks can vary but typically involves spoofing. This is where the adversary pretends to be a trusted source or authority, such as a bank or a well-known company, by forging email headers, URLs, or logos. They also use social engineering techniques to manipulate and deceive victims into providing sensitive information, such as usernames, passwords, or credit card details. This can be done through emails, websites, or phone calls (Venkatesha et al., 2021). At times they also masquerade as a colleague, friend, or family member to gain the victim's trust and convince them to share confidential information or perform certain actions.

Adversaries create a sense of urgency or fear, claiming that immediate action is required to avoid negative consequences, such as account closure, loss of funds, or legal actions. This urgency causes victims to click links that have malware (Jampen, et al., 2020). The malware can be part of phishing emails, text messages, or job post links. These are used to distribute malicious software, such as keyloggers, ransomware, or

remote access trojans (RATs), by enticing victims to download attachments or visit infected websites.

The victim's behavior when targeted by phishing attacks is usually due to falling for deceptions and believing the false information provided by the adversary, leading them to disclose sensitive information or perform actions they would not usually do (Jampen, et al., 2020). They end up clicking on the links in the communications the adversaries send them without verifying their authenticity or considering the potential risks and unknowingly provide personal details, such as passwords, social security numbers, or bank account information, to phishing attackers (Jampen, et al., 2020).

Many of the behaviors that victims engage in fall into the category of "lack to detail" where they ignore significant red flags. Warning signs or security measures, such as browser warnings about potential phishing attempts or untrusted websites can pop up but be ignored (Venkatesha et al., 2021). Furthermore, a lot of people who fall victim to phishing attacks fail to disclose it so that it may be examined and monitored. Failing to report phishing attempts or attacks to the appropriate authorities or IT departments could hamper efforts to prevent further incidents.

However, challenges persist in implementing behavioral approaches for phishing detection. One such challenge is the need for comprehensive data collection, as capturing various behavioral aspects without intruding on user privacy remains a concern. Additionally, attackers are continuously refining their techniques, which necessitates the development of adaptive detection algorithms. Despite these challenges, the evidence found provides effectiveness of behavioral approaches in phishing detection, emphasizing their potential in complementing existing solutions.

2.7 Solution to Phishing Attacks

Phishing attacks have become a major concern in today's digital world, targeting unsuspecting individuals and organizations to steal sensitive information such as usernames, passwords, credit card details, and financial information. However, there are several solutions available to combat these malicious activities.

Education and awareness play a crucial role in preventing phishing attacks. By providing proper training to individuals and organizations about the tactics used by phishers, users become more cautious and can identify phishing emails, messages, or websites. Common warning signs such as grammatical errors, suspicious URLs, and urgent requests for personal information should be shared with users to help them recognize potential phishing attempts.

Implementing robust security measures is another effective solution to phishing attacks. Companies need to invest in advanced email filtering systems that can detect and flag suspicious emails. Such systems use algorithms and artificial intelligence to identify phishing emails based on a range of factors, including email addresses, domain authenticity, and content analysis. Additionally, implementing two-factor authentication (2FA) can significantly reduce the risk of successful phishing attacks as it provides an additional layer of security by forcing users to provide an extra form of authentication credentials.

Regular software updates and patches also contribute to reducing the vulnerability to phishing attacks. Phishers often exploit security loopholes in outdated software to gain unauthorized access to systems. By keeping all software up to date, organizations are able to address known vulnerabilities, making it harder for phishers to exploit them.

Web browsers and email providers are continuously enhancing their capabilities to detect and block phishing attempts. For example, modern web browsers provide warnings and block access to suspicious websites that are known for phishing activities. Additionally, email providers automatically move suspicious emails to spam folders, preventing users from interacting with potentially harmful content.

Now these are all traditional methods of detecting and preventing phishing attacks and alone they are no longer sufficient as attackers are constantly evolving their tactics to evade detection. If these methods and machine learning are combined, a more robust solution against phishing attacks can be implemented. Machine Learning can offer a promising solution to combat phishing attacks. By harnessing the power of vast amounts of data and the ability to analyze patterns, machine learning algorithms can effectively detect and prevent phishing attempts in real-time. These algorithms can learn from previous phishing attacks, identifying common patterns and characteristics that differentiate a legitimate message or website from a malicious one.

One way machine learning can aid in phishing attack prevention is through the analysis of email content. Phishing emails often contain specific keywords, spellings, and grammatical errors that can be indicative of their malicious nature. Machine learning algorithms can be trained to identify these patterns and flag such emails as potentially dangerous, thereby minimizing the chances of users falling victim to phishing scams.

Machine learning can also be used to analyze website URLs and determine if they are legitimate or fake. Phishing websites often mimic the look and feel of legitimate websites, making it challenging for users to identify the deception. Machine learning algorithms can detect subtle differences in the structure and content of URLs, such as

misspellings or additional characters, that are common in phishing attacks. By promptly identifying and blocking such URLs, machine learning can help prevent users from accessing malicious websites.

Furthermore, machine learning can continuously adapt and improve its accuracy in detecting phishing attacks. As new phishing techniques are employed by cybercriminals, machine learning models can be trained using new data to recognize and adapt to these evolving threats. This process allows machine learning algorithms to stay one step ahead of attackers, making it increasingly difficult for them to deceive users (Sahingoz et al., 2019).

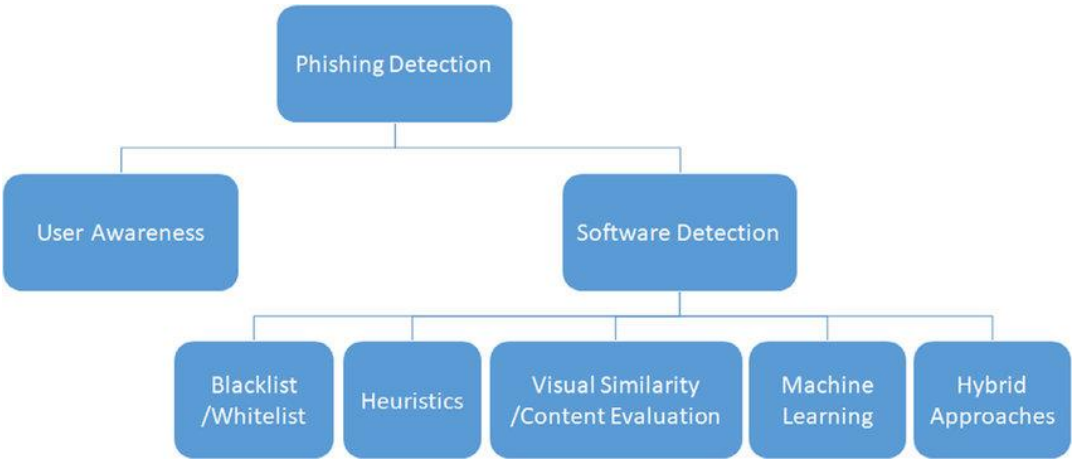


Figure 15: Solution to Phishing

In conclusion, machine learning provides a powerful solution to counter phishing attacks. By leveraging its ability to analyze patterns, machine learning can effectively detect phishing emails and websites in real-time. With the constant evolution of attack techniques, machine learning's adaptability helps in continuously strengthening phishing prevention measures. Integrating machine learning into security systems can significantly enhance cybersecurity measures, ultimately reducing the success rate of phishing attacks

and protecting users from falling victim to these malicious activities. Regular evaluation and improvement of these solutions are necessary to stay one step ahead of increasingly sophisticated phishing techniques.

2.8 Summary and Conclusion

Cyber-attacks cost economies millions if not billions of dollars each year where phishing is one of the most common attacks. Phishing attacks occur in many different forms. Attackers utilize a number of social engineering tactics and advanced security procedures to deceive their victims.

Several studies have highlighted the persuasive nature of social engineering techniques deployed in phishing attacks. Liu et al. (2017) examined phishing emails and identified common patterns in the manipulation of emotions, trust, and urgency to deceive users. Moreover, Campbell (2019) analyzed the psychological aspects of human vulnerability and emphasized the importance of user education to counter social engineering tactics.

While user education plays a critical role in mitigating risks associated with social engineering, additional measures are required to detect and prevent phishing attacks. For instance, recognizing patterns in the email content and sender behavior can aid in distinguishing legitimate communications from phishing attempts. Additionally, utilizing machine learning algorithms to analyze the linguistic features of phishing emails has shown promise in detecting social engineering-induced attacks. Previous studies show that deep learning, SVM, NLP and other neural networks were used to detect phishing attacks, mostly email or URLs.

The literature reviews in this chapter demonstrate the significance of social engineering techniques in phishing attacks, offering valuable insights for the development of robust detection mechanisms using machine learning and discusses the gap found in the data analysis due to data quality and validation in all the current research. The PhishNet is a system that will utilize data from all forms of phishing attacks to build a robust detection platform against phishing communication.

— Methodology

3.1 Introduction

As discussed in the literature review there are gaps in the phishing detection framework present in organizations worldwide. After researching it can be concluded that most phishing detection analysis is based on a blocklist of URLS, or machine learning models based on the anatomy of an email. In this praxis what has been utilized is a multifaceted and elaborate research design that integrates both the approach of quantitative and qualitative data collection and analysis methodologies. The quantitative component of this inquiry is slated to encompass the collection and analysis of data on phishing attacks through social engineering and user characteristics from an array of disparate sources, including incident reports and security logs.

On the other hand, the qualitative component of the study intends to undertake in-depth interviews with cybersecurity experts to glean their insights and perspectives on social engineering attacks and the efficacy of predictive models in detecting them. In essence, the mixed-methods research design is anticipated to provide an all-encompassing comprehension of the issue, aiding in identifying trends and patterns in phishing attacks through social engineering that can facilitate the formulation of predictive models to broaden the idea of one size fit all model based on one form of data – the URLs.

In this research journey, the primary objective was to construct a robust NLP-based phishing detection model. However, the path to achieving this goal was paved with a two-fold process. Firstly, the delving into the meticulous analysis of diverse datasets, each comprising different forms of phishing attacks. The aim was to gain comprehensive

insights into the characteristics and nuances of these attacks. By dissecting the standalone datasets, critical patterns, language structures, and behavioral triggers commonly employed by cyber adversaries, were unearthed. These insights laid the groundwork for a more profound understanding of the dynamics at play in phishing attempts, enabling us to identify the linguistic and contextual fingerprints that set them apart from benign communications.

The second crucial aspect of this research involved the integration of these datasets into a unified framework. The data was concatenated into a standalone dataset to create a rich and diverse repository that encapsulated a wide spectrum of phishing tactics. This amalgamation formed the basis for the machine learning endeavors. During the research a range of machine learning models were applied, each carefully tailored to leverage the unique characteristics extracted from the datasets. The primary objective was to create a sophisticated NLP-based phishing detection model that incorporated the knowledge and patterns identified during the analysis. The culmination of this two-pronged approach equipped us with a powerful tool capable of discerning even the most subtle indications of phishing in a sea of digital communication, contributing to a more resilient defense against these evolving threats.

The Natural Language Processing (NLP) model implemented was to identify and detect fake, scam, and fraudulent content present in various forms of communication, such as emails, messages, job postings, malicious URLs and other phishing postings by leverages the power of BERT (Bidirectional Encoder Representations from Transformers), a cutting-edge transformer-based model, that will aim to enhance the accuracy and robustness of the implemented phishing detection system.

To train and evaluate the phishing detection model effectively, a comprehensive dataset was curated consisting of five sub-datasets. These sub-datasets are specifically designed to cover a wide range of potential phishing scenarios, including phishing messages, phishing emails, phishing job postings, malicious URLs, and phishing GPT templates. The model is ensured to be well-equipped to handle numerous sorts of fraudulent information and to generalize effectively to real-world settings by merging these disparate data sources.

Before feeding the data into the NLP BERT model, the entire dataset was subjected to rigorous data preprocessing techniques. This step involved cleaning, tokenization, and handling missing values to ensure the data is in a suitable format for training. Additionally, Exploratory Data Analysis (EDA) was conducted to gain deeper insights into the distribution of data, the presence of class imbalances, and any potential challenges that might affect model performance. EDA allowed us to make informed decisions during model development and fine-tuning.

The core of the phishing detection system lies in the adoption of the state-of-the-art BERT model. To tailor the model for each dataset, an individual training and validation processes were carried out for each sub-dataset. By training the BERT model on each specific dataset, the aim was to understand the model's performance and adaptability to varying data distributions and class imbalances. Following this, a combination of the insights from all the individual validations and the two additional techniques utilized on the NLP model – the BART and XLNET - provided insight to draw meaningful conclusions about the overall performance and effectiveness of the BERT-based phishing detection system.

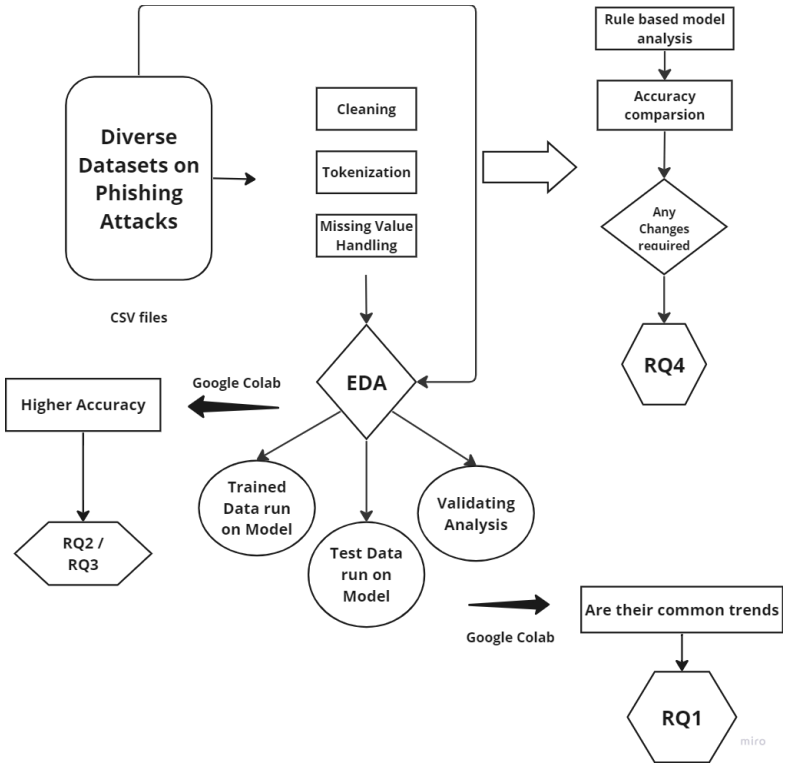


Figure 16: Praxis Research Methodology

3.2 Research Approach

The research is focused on using several datasets from a reliable source online known as Kaggle. A concatenation of the datasets is taken to create a Natural Language Processing (NLP) model using BERT for detecting phishing attacks. This is an innovative research approach that leverages the power of both crowd-sourced datasets and state-of-the-art NLP techniques. Kaggle, as a popular platform for data science competitions and datasets, offers a diverse collection of labeled and unlabeled data related to various domains, including cybersecurity. In this research attempt, the primary goal is to integrate and preprocess these datasets to train a BERT-based NLP model, thus enhancing the accuracy and robustness of phishing attack detection.

The first step in this research approach involves data collection and preprocessing. Multiple datasets from Kaggle that contain text-based information, such as email headers,

body content, or URLs, can be gathered. These datasets might include labeled examples of phishing emails, legitimate communications, or malicious URLs. Data preprocessing is critical, as it involves cleaning and harmonizing the data to ensure uniformity. Additionally, features like email addresses, URLs, and HTML tags may need special handling to ensure that BERT can effectively analyze the text.

Once the datasets are collected and preprocessed, the next step is to fine-tune a BERT model for phishing attack detection. Fine-tuning involves training the BERT model on the collected and preprocessed data to adapt it specifically to the task of identifying phishing attacks. During this process, the model learns to recognize linguistic patterns, contextual information, and semantic cues that are indicative of phishing attempts. Transfer learning, which allows the model to leverage its pre-trained knowledge on a vast corpus of text, is a key advantage of using BERT for NLP tasks like this.

The resulting BERT-based NLP model will then be evaluated using a number of the known performance metrics like accuracy, precision, recall, and F1-score. Additionally, it is essential to conduct rigorous testing with real-world phishing attacks to assess the model's effectiveness in a practical context. By combining Kaggle datasets and BERT-based NLP techniques, this research approach holds the promise of significantly improving phishing attack detection capabilities, thereby enhancing cybersecurity in an increasingly digitized world, and gaining a broader understanding of the various aspects related to phishing attacks.

3.3 Research Setting and Context

The subsequent sections of this report will outline the data used and the preprocessing process and methodology employed to have a balanced dataset. Through this analysis, the

aim is to provide an inclusive and comprehensive understanding of the efficacy of the NLP BERT model in identifying and mitigating potential phishing, frauds, and fake content. By assessing the model's performance on the diverse sub-datasets and analyzing its generalization capabilities, it is seen to offer valuable insights into the broader applicability of the phishing detection system in real-world scenarios. The goal is to contribute to the advancement of phishing detection techniques and promote a safe digital environment.

3.4 Data Collection Methodology

The comprehensive dataset used in this research consists of five sub-datasets all gathered from Kaggle. Kaggle is a known public data repository that provides different datasets on different topics for analysis purposes. These sub-datasets are specifically designed to cover a wide range of potential phishing scenarios, including phishing messages, malicious URLs, phishing emails, fake job postings, phishing recruitment-related content, and phishing templates. Given below is a brief overview of all these sub-datasets.

3.4.1 **Dataset 1: Phishing GPT Message Dataset**

This dataset consists of GPT templates of phishing messages used in cyber-attacks and is generated by different text-generated engines as well as social media platforms, online forums, and scam-reporting websites. It helps in identifying patterns and malicious elements used in phishing attempts. The number of instances in this dataset is 3792, consisting of both legit and phishing message templates. It consists of two columns, named as input text templates and a label column as "Phishing" or "Real" (1 or 0) entries respectively.

3.4.2 **Dataset 2: Phishing Text Messages Dataset**

This dataset comprises a collection of phishing text messages gathered from various sources, such as social media platforms, online forums, and scam-reporting websites. The messages encompass a wide range of scam scenarios, including lottery scams, phishing attempts, and fraudulent schemes. It consists of 5 columns, named “text”, “scam or not scam”, and 3 unnamed columns with almost all null objects in them. The number of instances/rows of this dataset are 5572.

3.4.3 **Dataset 3: Fake Job Posting Dataset**

This dataset includes records of job postings flagged as fake or misleading, typically promoting non-existent job opportunities or fraudulent hiring practices. It contains both categorical attributes as well as numerical attributes. The important attributes of this dataset are Job id, title, location, department, company profile, requirements, benefits, required experience, and fraudulent as labels. The total number of instances in this dataset is 17880 with some of the attributes having half or more than half null values in it.

3.4.4 **Dataset 4: Legit and Phishing Emails Dataset**

This dataset consists of emails flagged as phishing or identified as scam attempts by email security filters. The emails contain deceptive content, often promoting fraudulent products, services, or investment opportunities. Its columns/attributes are text and label, having a total 5172 number of instances in it.

3.4.5 **Dataset 5: Phishing URLs Dataset**

The phishing URL dataset is a collection of URLs that have been identified as potential phishing websites. These datasets are typically compiled by security researchers and organizations to study and analyze phishing attacks. They contain a wide range of

phishing URLs, including those that mimic popular websites like banking or social media platforms. The main attributes of this dataset are the URL and the type as labels. The total number of instances of this dataset is 651191 where there are 428103 benign, 96457 defacements, 94111 phishing and 32520 malicious URLS.

3.4.6 **Concatenation of Datasets**

The dataset attributes offer valuable insights into distinguishing legitimate content from deceptive and fraudulent information. This diverse data will form the foundation for the subsequent data pre-processing and Exploratory Data Analysis (EDA) stages, ensuring an effective and insightful model for detecting and mitigating fraudulent activities.

The new dataset created for this study is a comprehensive compilation of five distinct sub-datasets, each representing a unique facet of phishing-related scenarios. It consolidates phishing message template data, phishing text messages, phishing job postings, phishing emails, and phishing URLs, providing an extensive repository of fraudulent instances that can be used to study and model various scam behaviors.

The concatenation of these five datasets into a single unified dataset will enable a comprehensive analysis of phishing, facilitating the development of a robust and effective scam detection system. The data to create the concatenated dataset was split into equal samples of 50000 instances. These were further split into 10000 equal instances to test the hypothesis. The concatenated dataset had equal sample representations from standalone datasets.

3.5 Data Analysis Tool

The datasets are saved on csv files which provides us a number of ways to analyze the data through to better construct and validate the quality of the data. The following tools were utilized to analyze the data and make better judgement before finalizing it through Google Colab.

- **WEKA** is a common open-source data mining software that provides a collection of algorithms and tools that are mostly used when data preprocessing, classification, regression, clustering, association rules, and visualization is required to be evaluated on a given dataset.
- **Excel** is a spreadsheet program developed by Microsoft. Used for data analysis, data manipulation, and creating visualizations. It provides various functions, formulas, and features for organizing and analyzing data.
- **Minitab** is a statistical software package that is used for data analysis and quality improvement. It offers a range of statistical tools and techniques, such as hypothesis testing, regression analysis, control charts, and design of experiments.
- **Google Colab** is a cloud-based platform stipulated by Google for running and sharing Jupyter notebooks. It allows users to write and execute code in the Python programming language, collaborate with others, and access powerful machine learning libraries and frameworks. Google Colab provides free access to GPUs and TPUs.

3.6 Learning Algorithms Used

As previously mentioned, the goal of the praxis was to carry out an analysis that is data driven in order to implement a model that will help detect phishing language through various entry points. The final basic model is created using NLP BERT; however, a variety of machine learning techniques are utilized to support the analysis in order to implement a powerful, resilient model, obtain deeper insights, and expand the knowledge of the data used.Since the data quality has an impact on the predictive results an initial step is to start by analyzing the standalone datasets.

Each dataset was analyzed using one or more of the machine learning models that will be highlighted below. The outcomes of each independent analysis are compared to the final model analysis in Chapter 4. The first step was to implement the decision tree classifier which is a supervised learning algorithm that partitions the input space into mutually exclusive regions to make predictions. Second and furthermore, random forests, as an ensemble technique, amalgamates numerous decision trees to enhance precision while mitigating overfitting. Third, multilayer perceptrons (MLP), and one that took longer processing time than others, is a feedforward neural networks consisting of multiple layers of interconnected nodes that can capture complex patterns in the data.

Fourth, the XGBoost was utilized for some of the datasets which is a gradient boosting algorithm that creates an ensemble of weak learner models to build a more powerful predictive model. Fifth, and one of the weaker models, was the autoencoder neural networks which is unsupervised learning algorithms that can learn features from data by reconstructing the input from a compressed representation. The sixth step was the Bidirectional LSTM (Long Short-Term Memory) which is a type of recurrent neural

network (RNN) that processes in sequence data by considering past and future information. The seventh step was the Naive Bayes which is a probabilistic algorithm that applies Bayes' theorem with intense unconventionality expectations between features.

After utilizing these various learning algorithms to gain insights and extract additional features from the data, implementation of the NLP model using BERT (Bidirectional Encoder Representations from Transformers) commenced and the primary focus is on fine-tuning pre-trained BERT models for specific natural language processing tasks. While BERT itself is a powerful transformer-based model designed for a range of NLP tasks, fine-tuning is necessary to adapt it to particular applications. To help the model perform accurately, several libraries are installed within the Python code. One of which is Scikit-learn; this is one of the robust libraries in python that is commonly used in machine learning to stipulate several tools for statistical modeling like classification, regression, clustering, feature extraction and dimensionality. In the process of cleaning the data, it is split into X and Y axis – where X is the feature of phishing language and Y is the label of whether it is Legit (0) or Phishing (1). Then the Scikit-learn libraries are used to classify the data. Within the preprocessing phase of the data and identifying of data features that were picked up on during the standalone analysis and additional cleaning of text was run through the use of stop words and stemmers. Stop words are commonly used words like the, a, an, of, in and many others that are ignored as the entities are indexed and stemmer is a process where similarly spelled words are reduced like program, programming and programmer all are known as program.

It is important to note that BERT itself is a transformer-based language model that can be fine-tuned for a specific classification task. The classification algorithms

mentioned above are applied to the feature representations extracted from BERT, rather than directly to the BERT model itself.

Further experimentation with BART and XLNet was implemented. BART stands for Bidirectional and Auto-Regressive Transformers, which is a neural network architecture used for natural language processing tasks such as text classification and generation of texts. XLNet, on the other hand, is an autoregressive pretraining method that learns the probability of a sentence by training a generative model. This phase of the research was performed to assess the accuracy of the NLP BERT model. The concatenated dataset was run on both BART and XLNet. By comparing the results of these models with the performance of the NLP BERT model, evaluation of the effectiveness of the research's approach and determine if BART or XLNet could potentially provide higher accuracy or complement the existing model.

3.6.1 **Natural Language Processing**

NLP is part of the artificial intelligence field that focuses on understanding the interactions between humans and computers through its natural language. NLP helps computers to better understand, interpret, and then generate understandable human language by enabling machine language to process, analyze and generate the text/speech data. It is used to extract meaning from unstructured text data. This includes tasks like sentiment analysis, entity recognition, and topic modeling.

3.6.2 **Bidirectional Encoder Representations from Transformers**

BERT is a form of NLP. It is a deep learning model developed by Google AI based on the transformer architecture and is designed to understand the context and nuances of language by considering both the left and right context of each word in a sentence. BERT

is pre-trained on a massive corpus of textual language data and can be fine-tuned for a range of different NLP tasks. This model has significantly improved the accuracy of NLP tasks.

3.6.3 **Bidirectional and Auto-Regressive Transformers**

BART is a form of natural language processing model that combines bidirectional understanding of text (like BERT) with auto-regressive generation of text. This means it can analyze text in both directions, capturing context effectively, and it is proficient at generating coherent and contextually appropriate text. BART is particularly strong in tasks like text summarization, translation, and other text generation tasks where context and coherence are essential. It is a versatile model in the field of natural language processing.

3.6.4 **XLNet**

XLNet is a form of natural language processing model that is an extension of the Transformer architecture. It is known for its ability to model long-range dependencies in text. Unlike BERT, which reads text bidirectionally, XLNet considers all possible permutations of words in a sentence, making it more contextually aware. This helps it perform exceptionally well in a wide range of NLP tasks, including text classification, question answering, and language modeling. XLNet's versatility and its ability to capture complex relationships in text make it a powerful tool for understanding and generating natural language.

3.7 Model Architecture

In this subsection, the details of the base model's architecture used for the phishing detection are presented. The model architecture consists of the following components:

- **Tokenizer**: The tokenizer used for processing the input text is the BERT pretrained tokenizer. This tokenizer is specifically designed for BERT models and is responsible for converting raw text into numerical input that can be fed into the BERT model for processing.
- **TFBertModel**: The BERT model employed in this study is the TFBertModel with the architecture “bert base uncased”. This pretrained BERT model provides a powerful representation of text data by leveraging a transformer-based architecture with 12 layers and 768 hidden units.
- **Data Preparation**: To prepare the input feature values for the model, first pass the training data through the encoder function, which tokenizes the text and generates input IDs and attention masks. These datasets include text from emails, text messages, and bot messages. After combining them, the dataset contains 681,842 instances. The concatenated dataset had equal sample representations from each of the five standalone datasets.
- **Customized Build Model**: The Customized model takes the pre-trained BERT model as input and builds a classification model on top of it. A summarized Table 2 to represent this customized model architecture is given below:

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 128)]	0	[]
input_2 (InputLayer)	[(None, 128)]	0	[]
tf_bert_model (TFBertModel)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 128, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	109482240	['input_1[0][0]', 'input_2[0][0]']
dense (Dense)	(None, 64)	49216	['tf_bert_model[0][1]']
dropout_37 (Dropout)	(None, 64)	0	['dense[0][0]']
dense_1 (Dense)	(None, 1)	65	['dropout_37[0][0]']

Total params: 109531521 (417.83 MB)
Trainable params: 109531521 (417.83 MB)
Non-trainable params: 0 (0.00 Byte)

Table 2: Model Architecture

The core architecture is built using the BERT-based text classification model. It uses BERT's pre-trained embeddings and fine-tunes them for the specific phishing detection task. The model is set up to classify text instances into two classes: phishing or not phishing. The maximum sequence length for input text is limited to 350 tokens.

A learner object is created to manage the training process. This learner is initialized with the BERT-based text classification model, training data, and validation data. It plays a fundamental role in training and evaluating the model's performance.

The model is trained using a one-cycle learning rate policy with a learning rate of 2e-5. The training process occurs over 5 epochs, during which the model learns to associate input text data with their corresponding labels (phishing or not phishing). This process fine-tunes the BERT model specifically for the task of phishing content detection.

In summary, the architecture leverages the power of BERT, a pre-trained language model, to process and classify text data related to phishing content. It involves data preparation, data splitting, BERT-based preprocessing, model definition, and training. The objective is to create an effective model capable of distinguishing phishing language across a variety of text sources, including emails, text messages, and bot messages.

The model architecture table above summarizes the layers, output shapes, number of parameters (# Param), and the connections between each layer. The BERT model, which has 109,531,521 parameters, serves as the core component for text representation. On top of the BERT model, a Dense layer was added with ReLU activation, followed by a Dropout layer to prevent overfitting. Finally, the outcome is a Dense layer with a sigmoid activation for binary classification.

With this model architecture, the ability to effectively detect and mitigate fraudulent activities in various forms, such as phishing messages, phishing emails, fake job postings, phishing URLs, and phishing GPT templates is carried out. To measure the accuracy of the model a comparison with the two other pre-trained models within the NLP model – BART and XLNET is carried out and the results of these models will be discussed in Chapter 4.

3.8 Development of a New Framework

The work presented in this research provides us with enough evidence and the tools to improve the existing framework for detecting phishing attacks and being able to prevent them in both personal and business workspaces. Phishing attacks are a subset of cybersecurity attacks and the work of developing an improved framework connected to the existing frameworks to the MITRE Attack framework.

With the datasets used in the analysis, the works on identifying the Tactics is carried out, in addition to the objectives for attack and Techniques, methods used to achieve these attacks when phishing is performed. It is looked at how the victim's system was initially accessed. Most initial access was performed through social engineering tactics where the victim is tricked into clicking on a malicious link or downloading a virus through an attachment shared. Once gaining access is successful the adversary will then execute the malicious code and run the script to exploit victim’s vulnerabilities and collect information about the victim that include login credentials, financial data, or sensitive documents from their system which can disrupt operations or even damage reputations in some cases. What data was gained from the access will help provide us a better understanding of the objectives behind the attacks.

Prevention and detection of cyber-attacks is an important layer of the MITRE attack framework that extended on in the framework to help organizations enhance their ability to prevent phishing attacks and understand the different techniques that adversaries utilize to gain access to systems illegally by implementing the appropriate security measures. The phishing model which was developed to identify phishing language based on its usage and training on phishing data from various sources was used as a base. This helps security teams understand the variations of phishing attack tactics and tailor the countermeasures that best fit their organization against these attacks.

The framework provides a structured approach to mitigating the tactics and techniques of phishing attacks performed by adversaries which will help organizations enhance their overall cybersecurity postures and readiness against these attacks and the

ability to share threat intelligence information that can help identify phishing campaigns and provide the indicators of compromise.

Phishing Objectives	Countermeasures
INITIAL ACCESS	Preventing access through the use of PhishNet model that examines the language of various text formats and blocks malicous messages from getting accessed.
EXECUTION	Prevent any malicious link or file from being clicked or downloaded by blocking unrelated business code from running and creating an attachment check system.
DEFENSE EVASION	Understanding urgency level and impersonation techniques that adversaries utilize to carry out their attacks we would add a layer of internal communication to build a defense against threats.
CREDENTIAL ACCESS	Create multi-level authentication to prevent credentials from being accessed.
DISCOVERY	As part of a future research this can be implemented using the browser extension that will block any malicous message. This extension can be implemented on computers, iPads and phones.
LATERAL MOVEMENT	Security system alert when an account without access is accessing sensitive information.
COMMAND AND CONTROL	Identifying the various techniques used to perform phishing attacks commands to control that would be coming from an adversary is identified and block to pervent command and control.
IMPACT	If an attack occures understanding the impact of it and what damage it produces gives security teams a faster response time against these attacks.

Table 3: Phishing Security Controls Developed Through the MITRE Attack Framework

The MITRE Attack Framework is based on providing security teams with information about tactics, techniques, and procedures – the TTPs used by the adversary during different stages of the cyberattack. The phishing detection framework that was developed follows in the same footsteps focusing on the methods, tools, and the strategies required to identify and mitigate phishing attacks and provide the security team with a new way to defend against social engineering and fraudulent communication. This is done through focusing on the human element that falls victim to deceptive communication to enhance the organization's ability to recognize and respond faster to phishing attacks.

3.9 Conclusion

In summary, this research was part of the novelty of expanding the one approach to detecting phishing attacks. The gap that was discussed in the previous chapters, shows

that all phishing detection models were either based on one feature check: URL block list, Email text or Message text. By creating a fit all phishing attack methodology there were five hypotheses that had to be tested and helped produce an NLP model that will detect phishing attacks that could be part of any fraudulent text or speech to text data. The next chapter will dive into the results of the generated methodology and show the difference of one approach model versus the concatenation approach model.

— Results

4.1 Introduction

In today's digital age, the prevalence of online communication has made individuals and organizations more susceptible to cyber threats, particularly phishing attacks. Phishing, a malicious practice where attackers deceive individuals into revealing sensitive information or performing harmful actions, has evolved into a sophisticated and pervasive threat. Phishing attempts often involve manipulating language to exploit human emotions, fears, and urgencies, making them harder to detect.

The primary goal of this research endeavor is to harness the potential of natural language processing models and machine learning techniques in crafting a predictive model tailored for the strategic detection of phishing language across various communication channels. To guide this pioneering exploration, a formulated a series of research questions was conducted:

R1: **Leveraging Diverse Datasets**: To enhance the model's adaptability to evolving threats, does working with extensive and diversified datasets significantly boost the NLP model's capacity to detect emerging and novel phishing attack strategies?

R2: **Assessing Content Credibility**: How effectively can an NLP-driven model evaluate the credibility of email, message, or text content to thwart phishing attempts?

R3: **Comparing NLP with Rule-Based Approaches**: In a head-to-head evaluation, how does the performance of an NLP-based phishing detection model measure up against traditional rule-based techniques?

These research questions have been substantiated with corresponding hypotheses:

H1: **Adaptability to Novel Tactics**: An NLP model, trained on a vast and diverse dataset, has the potential to swiftly adapt to emerging phishing tactics and patterns, resulting in enhanced detection rates and the prevention of novel attacks.

H2**: Content Credibility Assessment**: By scrutinizing the linguistic and contextual attributes of email, text, and message content through natural language processing techniques, the model can accurately gauge content credibility, fortifying its ability to thwart phishing attacks.

H3: **NLP vs. Rule-Based Comparison**: The NLP model, enriched with contextual information and semantic analysis, can outperform conventional rule-based approaches, excelling in terms of accuracy and adaptability while detecting and preventing phishing attacks.

To achieve this, it is imperative to address the crucial problem of spotting phishing efforts in a variety of digital communication channels, including job posts, emails, malicious URLs, and text messages. The objective is to build a strong model that can identify possible phishing material from a variety of datasets by utilizing Natural Language Processing (NLP) techniques, particularly BERT (Bidirectional Encoder Representations from Transformers) and other machine learning approaches.

The datasets utilized for this project include:

- **Fake Job Posting Dataset**: This dataset contains examples of deceptive job postings, a common phishing tactic aimed at job seekers.

- **Malicious and Benign URL Dataset**: URL analysis is crucial for identifying phishing websites.
- **Phishing GPT Message Dataset**: This dataset includes text messages generated by malicious bots, which can contain phishing content.
- **Phishing Text Message Dataset**: A dataset containing text messages used for phishing attempts.
- **Legit and Phishing Emails Dataset**: Email is a prominent medium for phishing attacks, and this dataset captures examples of both legitimate and phishing emails.

By concatenating these datasets and employing a state-of-the-art NLP BERT model, the goal is to develop a powerful tool for differentiating between phishing and non-phishing content across these various digital channels. The concatenated dataset had equal sample representations from each of the five standalone datasets.

Furthermore, this project intends to go beyond binary classification by conducting a thorough feature analysis of the language used in phishing content. This analysis will explore linguistic patterns associated with urgency, fear, and financial transactions, among other factors. Understanding these language trends and ranking their significance will provide valuable insights into the characteristics of phishing attempts.

The results below showcase the outcome resulted during the analysis of the five phishing datasets in addition to the concatenated dataset. It shows a statistical analysis of the quality of data used in the analysis for each dataset separately and as concatenation dataset through data concatenation process. The data quality improvement is showcased in addition to identifying the data quality improvement methods, and how the data was

used to build the NLP phishing model. The findings and outcomes of this study, rooted in the prescribed methodological framework, are comprehensively presented in this chapter.

4.2 Dataset Quality Metrics

High-quality datasets are the bedrock of effective machine learning models and data-driven research. To ensure the trustworthiness and utility of a dataset, various metrics are employed to assess its quality. These metrics encompass several critical aspects, including data completeness, accuracy, and consistency. Completeness measures the extent to which the dataset covers the intended scope of the research which in the project is covering phishing attacks from different communication vectors and highlighting the presence of any missing or incomplete data points. Accuracy gauges the precision of the data, ensuring that it faithfully reflects the real-world information in presenting phishing data. Consistency will check for uniformity in data formatting, naming conventions, and values, preventing discrepancies that might hinder analysis.

The importance of dataset quality metrics cannot be overstated. This process serves as the gatekeeper for ensuring the reliability and effectiveness of the machine learning model and research outcomes. High-quality datasets enhance the robustness and generalizability of models, enabling them to make sound predictions and inferences. In research, the integrity of findings is heavily reliant on the quality of the underlying data. Biased, incomplete, or inaccurate datasets can lead to misleading conclusions and hinder the validity of research results. The meticulous evaluation and maintenance of dataset quality metrics are pivotal in advancing the credibility and impact of the machine learning and data-driven research.

4.2.1 **Data Feature Engineering**

Data feature engineering is the process of modifying or creating new features from raw data to improve and progress the performance of machine learning models. It involves selecting, transforming, and combining variables to extract relevant information that can enhance model accuracy and predictive power. This technique aims to uncover patterns, relationships, or insights hidden within the data, which might not be initially apparent.

The importance of feature engineering lies in its ability to significantly impact the accuracy and effectiveness of machine learning models. Well-engineered features enable models to capture intricate relationships and hidden patterns, making them more powerful and capable of generalizing to new data. It can lead to faster model training, improved interpretability, and reduced overfitting. Feature engineering can also help in reducing the dimensionality of the dataset, making it computationally more efficient and potentially revealing the most salient aspects of the data.

In the research, keyword extraction is used by extracting all relevant words that focus on words that convey urgency, fear, or deception using techniques such as tokenization to breakdown the text. Then the features were classified by creating a binary frequency feature for each feature to define the frequency of its presence in each of the phishing datasets. The six categories extracted were words featuring urgency, financial desire, curiosity, fear, job requirement and entertainment purposes. The main factor that prevented victims from falling for phishing attempts was shown to be urgency.

A closer look at each of these characteristics and a calculation of their accuracy using machine learning classifiers allowed researchers to validate the potency of the language

used to trick people into falling for phishing scams. Below a showcase of the overall feature extraction and then the classification of the features through five different techniques – Decision Tree Classifier, Random Tree Classifier, Multilayer Perceptron's (MLPs), XGBoost Classifier and Autoencoder Neural Network.

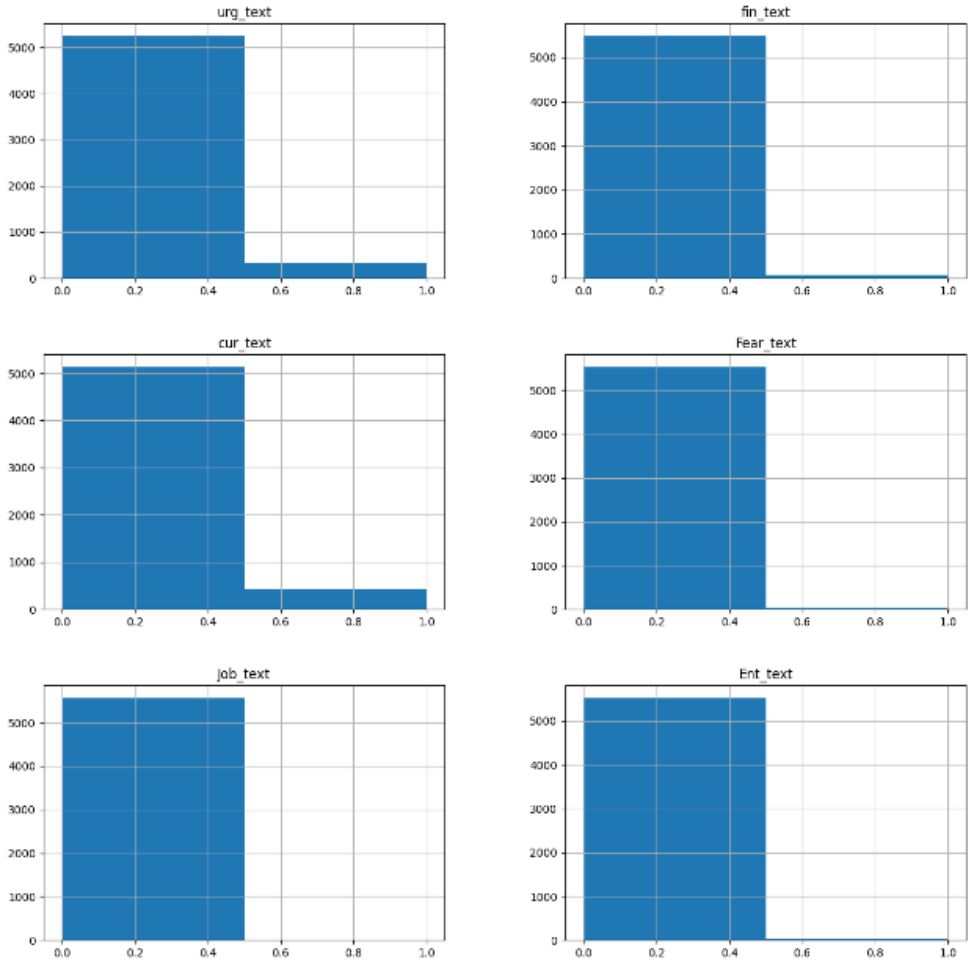


Figure 17: Data Distribution of the Six Features Present in Phishing Datasets

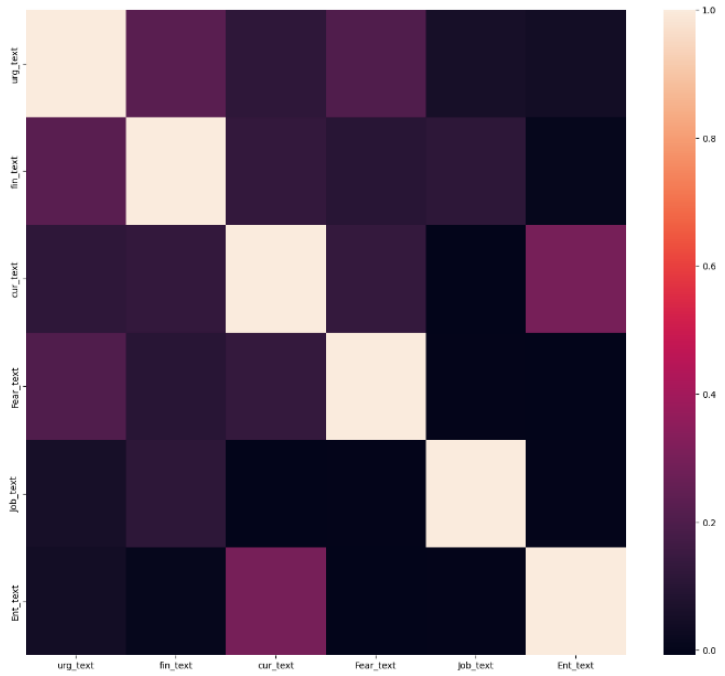


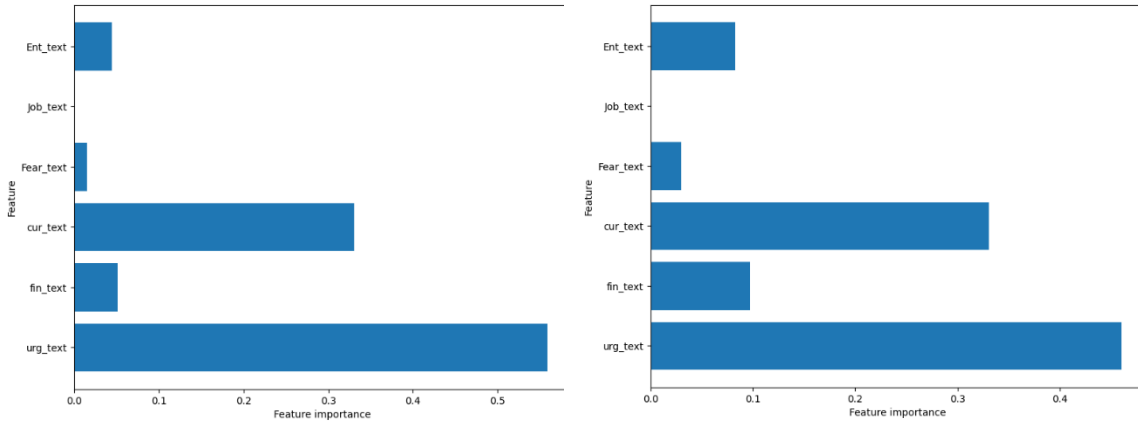
Figure 18: Correlation Heatmap of the Six Features

Feature	Urgency	Financial Desire	Curiosity	Fear	Job Requirement	Entertainment
Mean	0.057609	0.014537	0.077531	0.006820	0.000179	0.008794
Standard Deviation	0.233024	0.119701	0.267455	0.082307	0.013397	0.093371

Table 4: Average of Each Feature in Phishing Data

To further analyze the features of the dataset, the dataset is put under a number of tests to confirm the reliability of the feature engineering process. First, the Decision Tree Classifier is used where it systematically partitions the data into subsets based on the phishing data features and their values, which allowed the classifier to create a tree-like structure of decisions that ultimately leads to classification. This helps identify and confirm the key characteristics that most effectively distinguish between legitimate and

phishing messages, aiding in accurate detection. Second, the Random Forest classifier is used which is an ensemble machine learning method that instead of running on one decision tree uses multiple decision trees to analyze data. It exceled in measuring the phishing data features by aggregating the results of numerous decision trees to identify the most influential features for accurate classification, offering robust and precise detection capabilities. Under both classifiers it is clear that urgency ranks the highest making that the number one reason people fall victims to phishing attacks.



(a) Decision Tree Classification (b) Random Forest Classification

Figure 19: Decision Tree vs Random Forest Classification of Phishing Data Features

Three more machine learning algorithms are used in an effort to measure phishing data attributes and the accuracy of the data being phishing vs legitimate. The MLP classifier effectively extracts and processes the phishing features to capture intricate relationships within the data, contributing to the model's capacity to accurately distinguish between phishing and legitimate content. While what XGBoost performed was to systematically evaluated the importance of each feature, assigning weights to them and identified the most influential attributes that aid in precise and efficient phishing

detection. And the last neural network used was the autoencoder which is initially designed for unsupervised learning that learns to encode and decode data efficiently. In the data, it was able to reduce the dimensionality of the data while preserving its critical information, facilitating the identification of key attributes for phishing detection. The accuracy for all five classifiers is distributed below with Autoencoder providing the highest accuracy level.

ML Model	Train Accuracy	Test Accuracy
AutoEncoder	0.982	0.983
Random Forest	0.889	0.900
Multilayer Perceptron's	0.889	0.900
XGBoost	0.889	0.900
Decision Tree	0.889	0.899

Table 5: Compare the Models Performance on Feature Engineering

The key to success in feature engineering for phishing detection is to have a deep understanding of the characteristics of phishing messages and the ability to translate these patterns into actionable features. It is essential for improving model performance, addressing data quality issues, leveraging domain knowledge, and enhancing interpretability. By extracting and engineering features that represent urgency, fear, and deception, machine learning models can empower the model to identify phishing attempts effectively, regardless of the communication medium.

4.2.2 **Data Pre-Processing and Exploratory Data Analysis**

To obtain quality of data that will provide valid accuracy to the model a process known as EDA – Exploratory Data Analysis is performed. This step is essential in the data analysis process. It involves visually and statistically summarizing, examining, and understanding the main characteristics, patterns, and relationships within a dataset. This step helped us gain insights, identify anomalies, and formulate hypotheses to make

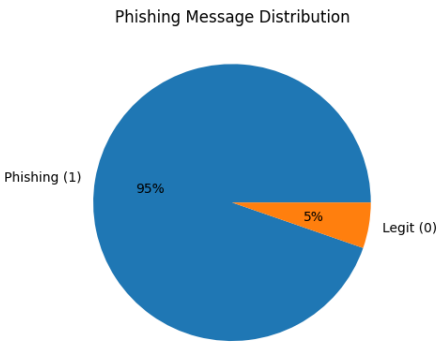
informed decisions about the subsequent data modeling, hypothesis testing, and machine learning tasks that are to be taken to build the phishing detection model.

When running EDA on the five chosen datasets, the ability to identify missing values, inconsistent data, or errors that need to be addressed before further analysis is clearer. This helped understand the nature and structure of the dataset as it involves creating visual representations of data, such as histograms, box plots, and scatter plots, frequency tables to aid in understanding and communication.

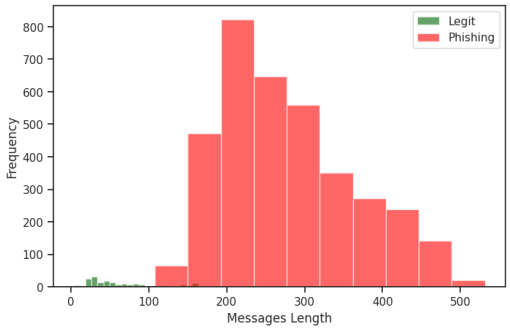
An overview of the types of characteristics and the number of occurrences of each dataset is provided in the detailed descriptions of each dataset found in Chapter 3. In this subsection, the pre-processing techniques, Exploratory Data Analysis to observe the insights of given datasets, and post-processing techniques applied to each dataset are explained and briefed.

In the phishing GPT message dataset, the attributes with considerable null values are eliminated for data accuracy. The label attribute contains 3590 instances of class “Phishing” while 202 instances of class “Legit” as shown in Figure 20 (a).

A new attribute “length” was added to the dataset, representing the length of each text feature instance to observe the lengths of the messages of each class label as shown in Figure 20 (b). Text preprocessing techniques, such as cleaning, removing stopwords, and stemming, were applied to post-process phishing email templates using the “stopwords” and “stemmers” packages of the Natural language tool kit.



(a) Input features classification distribution



(b) Message features length Histogram

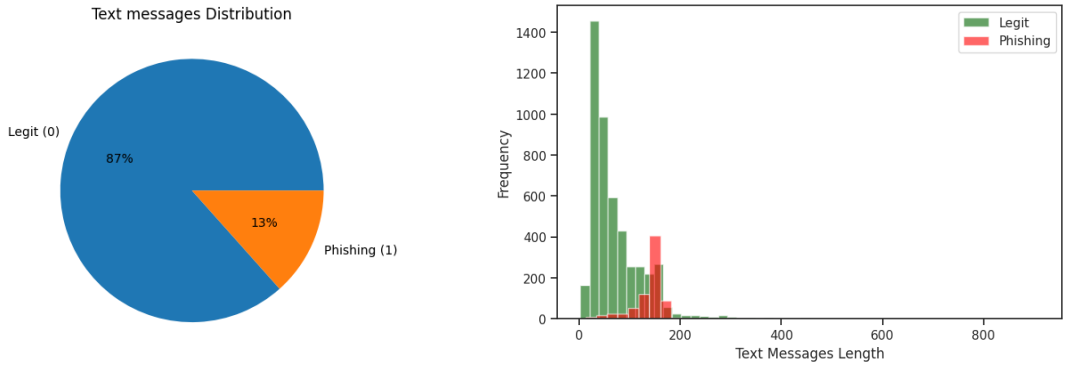
Figure 20: Exploratory Data Analysis for Dataset-1- Phishing GPT Messages

Phishing Message Data	
Mean	0.946729958
Standard Error	0.003647352
Median	1
Mode	1
Standard Deviation	0.224601086
Sample Variance	0.050445648
Kurtosis	13.84837975
Skewness	-3.980084928
Range	1
Minimum	0
Maximum	1
Sum	3590
Count	3792
Confidence Level (95.0%)	0.007150962

Table 6: Descriptive Analysis of Phishing GPT Message Dataset

In the phishing text message dataset, the three unnamed attributes with considerable null values are dropped for data accuracy. The input text features, and label columns were renamed. Out of 5572 total instances in this dataset, 87% instances belong to the class label “Legit”, and the remaining 13% belong to the class label “Phishing” as shown in Figure 21 (a).

A new attribute “length” was added to the dataset, representing the length of each text feature instance to observe the lengths of input feature instances of each class label as shown in Figure 21(b). Text preprocessing techniques, such as cleaning, removing stopwords, and stemming, were applied to post-process scam messages input features using the “stopwords” and “stemmers” packages of the Natural language tool kit.



(a) Input features classification distribution

(b) Message features length Histogram

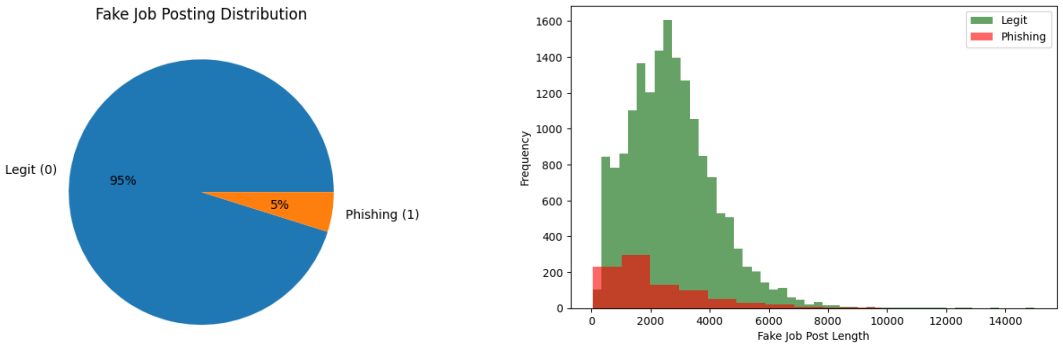
Figure 21: Exploratory Data Analysis for Dataset-2- Phishing Text Messages

Phishing Text Message	
Mean	0.134063173
Standard Error	0.0045649
Median	0
Mode	0
Standard Deviation	0.340750755
Sample Variance	0.116111077
Kurtosis	2.617413415
Skewness	2.148598161
Range	1
Minimum	0
Maximum	1
Sum	747
Count	5572
Confidence Level (95.0%)	0.008948984

Table 7: Descriptive Analysis of Phishing Text Message Dataset

In job posting dataset, the undesired attribute columns i.e., “job id”, “telecommuting”, “has company logo”, and “has questions”, with considerable null values are dropped for better visualization and model accuracy. Two new attribute columns, “title” and “department” were generated by merging the job description relevant attributes and attribute columns containing the department description respectively. The attribute values with null values were replaced with “No data” values and these input feature attribute columns contain a total of 17880 instances. The label class distributions are as shown in Figure 22 (a).

A new attribute “length” was added to the dataset, representing the length of each text feature instance to observe the lengths of input feature instances of each class label as shown in Figure 22 (b). Text preprocessing techniques, such as cleaning, removing stopwords, and stemming, were applied to post-process text input features using the “stopwords” and “stemmers” packages of the Natural language tool kit.



(a) Input features classification distribution (b) Message features length Histogram

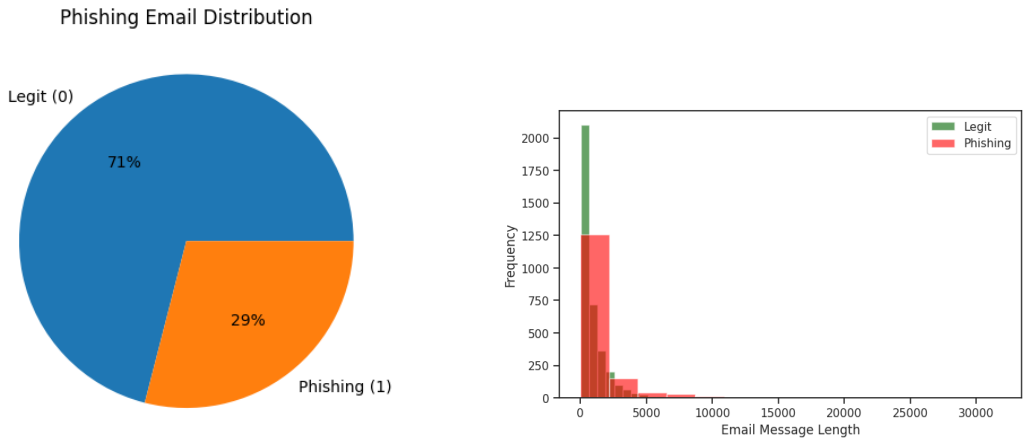
Figure 22: Exploratory Data Analysis for Dataset-3: Fake Job Posting

Fake Job Posting	
Mean	0.048434
Standard Error	0.001606
Median	0
Mode	0
Standard Deviation	0.214688
Sample Variance	0.046091
Kurtosis	15.70228
Skewness	4.207199
Range	1
Minimum	0
Maximum	1
Sum	866
Count	17880
Confidence Level (95.0%)	0.003147

Table 8: Descriptive Analysis of: Fake Job Posting Dataset

In phishing email dataset, out of the initial two columns, “text” and “label”, the text column has the legit email bodies. The columns are renamed and reordered by the names of Input text “features” and “label”. The input feature attribute column contains a total of 5172 instances with 1500 instances belonging to class “Phishing” and the remaining 3672 instances belonging to class “Legit”. The label class distributions are shown in Figure 23 (a).

A new attribute “length” was added to the dataset, representing the length of each text feature instance to observe the lengths of input feature instances of each class label as shown in Figure 23(b). Text pre-processing techniques, such as cleaning, removing stopwords, and stemming, were applied to post-process text input features using the “stopwords” and “stemmers” packages of the Natural language tool kit.



(a) Input features classification distribution (b) Message features length Histogram

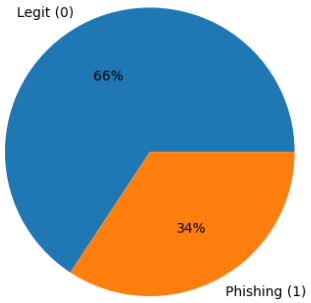
Figure 23: Exploratory Data Analysis for Dataset-4 – Phishing Emails

Phishing Emails	
Mean	0. 290023
Standard Error	0.001606
Median	0
Mode	0
Standard Deviation	0.453817
Sample Variance	0.046091

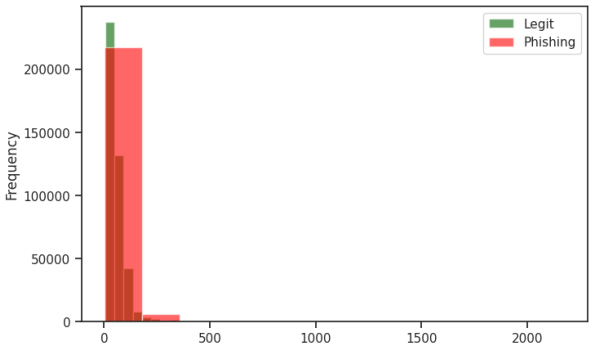
Table 9: Descriptive Analysis of: Phishing Emails Dataset

The Malicious URL dataset initially has 2 attribute columns of which one is the URL and the other is a categorical category. All the redundant and undesired attribute values were dropped. The attribute columns containing null values were replaced by “No Data” values and the correlated attributes were dropped and some relevant attributes were merged to form the final input feature column, named “URL” and the fraudulent column is named “type”. All the rows that have empty values in input feature column were also dropped. The processed dataset has total of 651191 instances with 428103 instances of class “Legit” and 223088 instances of class “Phishing” as shown in Figure 24 (a)

A new attribute “length” was added to the dataset, representing the length of each text feature instance to observe the lengths of input feature instances of each class label as shown in Figure 24 (b). Text pre-processing techniques, such as cleaning, removing stopwords, and stemming, were applied to post-process text input features using the “stopwords” and “stemmers” packages of the Natural language tool kit.



(a) Input features classification distribution



(b) Message features length Histogram

Figure 24: Exploratory Data Analysis for Dataset-5 – Malicious URL

Malicious URL	
Mean	0.342584587
Standard Error	0.000588098
Median	0
Mode	0
Standard Deviation	0.474574266
Sample Variance	0.225220734
Kurtosis	-1.559907399
Skewness	0.663398366
Range	1
Minimum	0
Maximum	1
Sum	223088
Count	651191
Largest(1)	1
Smallest(1)	0
Confidence Level (95.0%)	0.001152654

Table 10: Descriptive Analysis of: Malicious URL Dataset

The concatenation of the five datasets into a new one necessitated a comprehensive Data Pre-processing and Exploratory Data Analysis (EDA) process to ensure data accuracy, ease of visualization, and robustness of the subsequent data models. The concatenated dataset had equal sample representations from each of the five standalone datasets.

For the phishing message dataset, null values were eliminated, and a new “length” attribute was introduced. The label attribute showed a larger proportion of “Phishing” instances compared to “Legit”. After pre-processing, the data was cleaned, stopwords removed, and stemming applied using the “stopwords” and “stemmers” packages of the Natural language tool kit.

The phishing messages dataset underwent a similar pre-processing where the unnamed attributes with null values were dropped. The input text features, and label

columns were renamed, and the addition of a “length” attribute allowed for further insights into the distribution of message lengths across “Legit” and “Phishing” class labels.

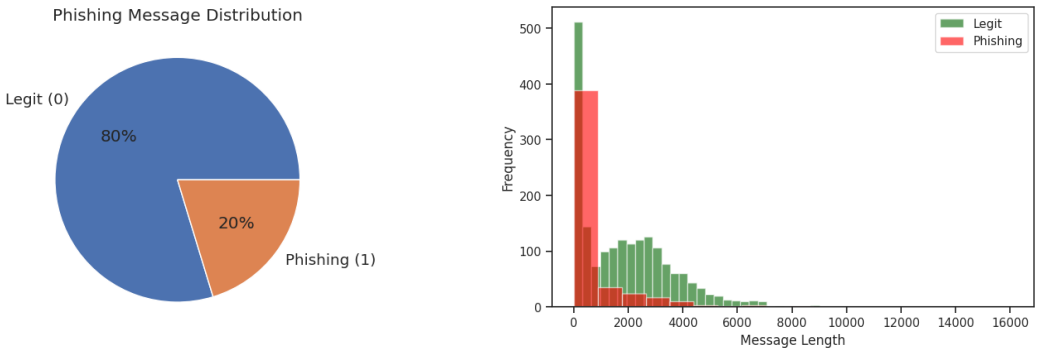
For the fake job posting dataset, attributes with substantial null values were dropped and the remaining data was enhanced by generating new columns, “title” and “department”. Null values were replaced with “No data”, and text preprocessing techniques were applied to the newly formed input feature attribute columns.

In the Phishing emails dataset, the columns were renamed and reordered. A “length” attribute was introduced to observe the lengths of input feature instances of each class label. Text pre-processing techniques were applied to the input text features to ensure data cleanliness and relevance.

Finally, the Malicious URL dataset underwent substantial modifications. Redundant attribute columns were dropped, null values replaced, and relevant attributes were merged to form the “URL” input feature column. The final dataset comprised 651191 instances with a significant majority belonging to the “Legit” class.

Overall, this extensive pre-processing and exploratory data analysis of each dataset was instrumental in ensuring the robustness and integrity of the unified dataset, providing a solid foundation for subsequent analyses and modeling. The combined data consisted of 683607 entities where 229044 where phishing and 448991 were legit and had equal sample representations from each of the five standalone datasets.

To run the analysis and be able to get the results required a split of the data into a 50000 sample is performed and split that into both a training, testing and validation dataset.



(a) Input features classification distribution

(b) Message features length Histogram

Figure 25: Exploratory Data Analysis for Concatenation of Datasets

4.2.3 Identifying and Rectifying Data Quality Issues in Training Data

As discussed in the previous section, the preprocessing process has to take place first for the datasets to be used for training purposes. This is done to ensure that the metadata used for the research that will run on both WEKA and Google Colab will match between both the training and testing datasets. As seen that some of the datasets required some minor tweaking to obtain true values to provide a true random split between the training and testing datasets.

Identifying and rectifying data quality issues in training data is crucial to ensure the accuracy and reliability of machine learning models. Training data serves as the foundation for the model, and any deficiencies or biases in the data can greatly impact the model's performance. Poor data quality can severely impact the performance and reliability of the model and can lead to erroneous conclusions.

Within the data analysis of the data, ability to identify the missing data. Missing values can distort statistical analysis and modeling. The use of heatmaps helped visualize the areas where data missing could affect the model. Within some datasets rows were removed with missing values if they are a small proportion of the dataset and do not carry critical information or imputed the missing values using statistical techniques like mean, median, mode imputation or more advanced methods like regression imputation.

When working on the data preprocessing step it is very important to identify all unusual data points. Like outliers, these deviate significantly from the rest of the data. They can affect model training by skewing statistics or introducing noise. Duplicate fields as well are another item that required cleaning up. This ensures that the training data is reliable and representative, which is essential for building accurate and robust a machine learning model and conducting meaningful data analysis. Data quality checks are an ongoing process for us to maintain the integrity of the dataset as new data is collected or as the dataset evolves over time.

4.3 Data Assessment and Evaluation

Dataset assessment and evaluation in the context of developing the Natural Language Processing (NLP) model with emphasis on BERT is a critical phase of the research as it ensures the reliability, effectiveness, and ethical use of the machine learning model. This process involves a comprehensive examination of the dataset's quality, balance, and representativeness, followed by rigorous model evaluation to gauge its performance.

When working with phishing datasets, dataset assessment and model evaluation are crucial due to the deceptive and harmful nature of phishing attacks. A poorly assessed or evaluated dataset can lead to false security, allowing phishing messages to go undetected,

or excessive false alarms, causing undue disruption. By rigorously assessing and evaluating datasets and models, you can build reliable and effective defenses against phishing attacks while maintaining ethical and privacy standards. Every dataset is observed and evaluated as a standalone dataset and then as a combined dataset as it is trained on the PhishNet model. The data used is run through a complete data validation process before being evaluated by the NLP -BERT model to help test the hypothesis.

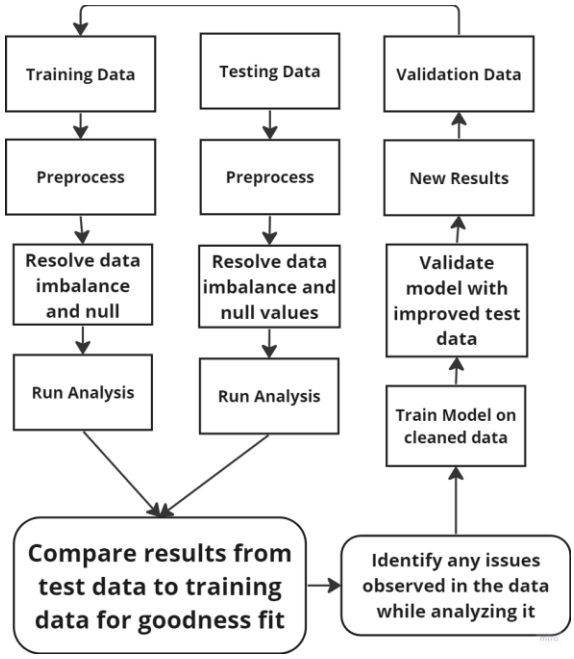


Figure 26: Data Validation Process

4.3.1 **Standalone Datasets Assessment**

Assessing each dataset as a standalone entity before integrating it into the model is crucial for several reasons. Each dataset might come with its unique set of characteristics, quality, and biases. By thoroughly evaluating each dataset independently, the ability to identify and rectify potential issues, such as missing data, inconsistencies, or imbalances, which could impact the model's performance is done. Moreover, it allows us to understand the specific context and domain that each dataset represents, tailoring the

model's preprocessing and training strategies accordingly. This individual assessment ensures that the harness of the full potential of each dataset while mitigating the risk of introducing noise or unintended biases into the model, ultimately resulting in a more robust and reliable NLP model for detecting phishing attacks across various scenarios.

- **Dataset 1:** Phishing GPT messages Dataset: With a dataset of 3792 GPT messages of both phishing and legit language to confirm the accuracy of the data where 5.33% of it was phishing, the accuracy was tested using gaussian where it was 98% and multinomial where it was 99% accurate, then with Bernoulli and 99.6% accuracy was retrieved. These metrics provide us with the confidence that when the model is trained on the data, accurate detection of phishing attacks is carried out.
- **Dataset 2:** Phishing Text Messages Dataset: In the phishing dataset there were 5163 text messages that were categorized between phishing or legit where 12.53% of the data was phishing texts. Using a number of ML techniques like PorterStemmer and Stopwords the text was transformed then used a number of algorithms to measure the accuracy of the data after transforming the text to confirm if the correct categorization of phishing and legit was being outputted. Similar to the GPT messages the accuracy was calculated using gaussian where it was 86% and multinomial where it was 98% accurate, then with Bernoulli and 97% accuracy was retrieved. Further examining of the phishing text on additional classifiers was run to further confirm the scores outputted in the analysis.

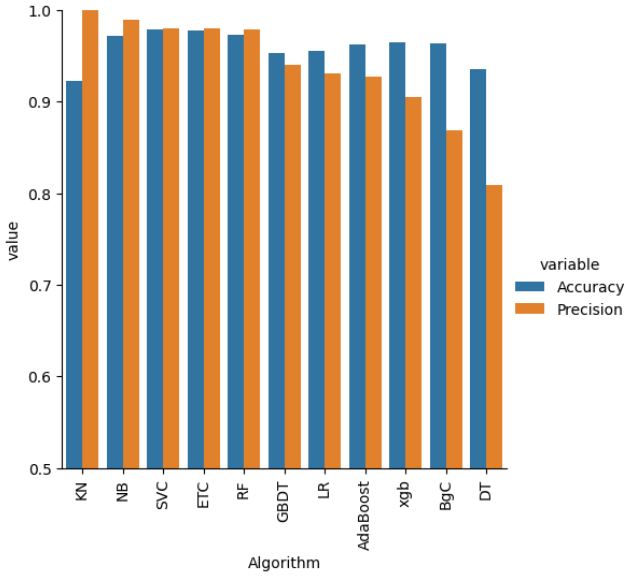


Figure 27: Phishing Text Accuracy and Precision of its Categorization

- **Dataset 3:** Fake Job Posting Dataset: This dataset as discussed previously is a mix of phishing and legit job posts consisting of 17888 posts. About 10% of the postings were phishing. Job postings are more difficult to analyze as there are a number of different features that affect the credibility of the posting. So, to validate the accuracy of this data set a base NLP model was used with Early stopping and Model checkpoint and got an accuracy of 98%.
- **Dataset 4:** Phishing Emails Dataset: Out of 3423 emails they were categorized as phishing or legit emails. To confirm the accuracy of the dataset and the validation of these emails as phishing or not they were classified via Random Forest and Multinominal Classifiers. The accuracy for Multinomial Classifier was 96.99 % and the accuracy for Random forest Classifier: 97.52%. In addition, phishing emails were analyzed to be longer than legit emails.

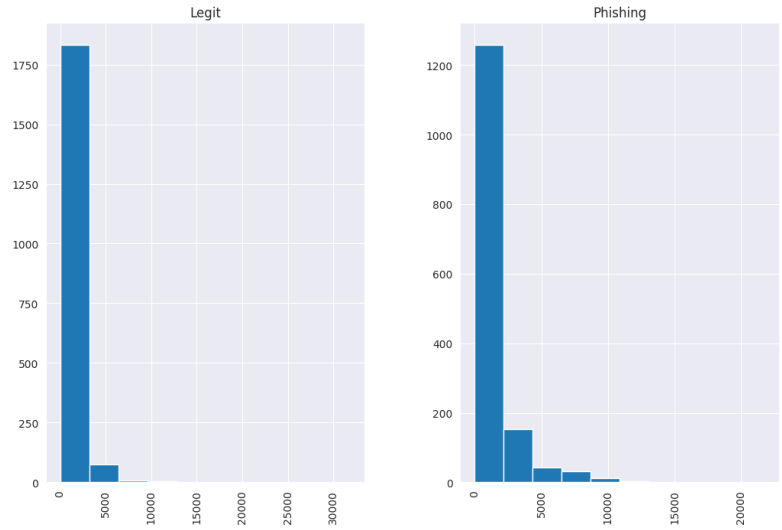


Figure 28: Distribution of Phishing and Legit Email Length

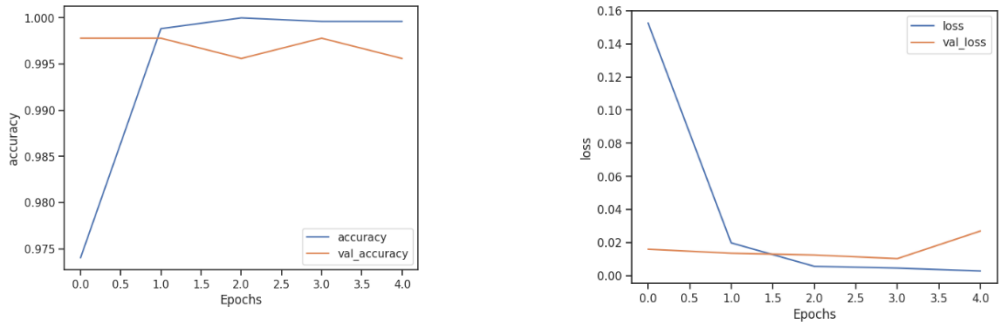
- **Dataset 5:** Malicious URL Dataset: The largest of all datasets since this is the most common dataset used for phishing detection in most organizations. The dataset consists of 651191 URLs each categorized as phishing or legit. The phishing URLs also fall within malware and defacement URLs. Feature engineering was performed to define the importance of each against it being phishing or not. Multilayer Perceptron's accuracy was 0.966, Random forest accuracy was 0.908 and Decision Tree accuracy was 0.906.

4.3.2 Standalone DataSets Evaluation on PhishNet Model

Evaluating the NLP-BERT model on each dataset as a standalone entity is of paramount importance as it ensures the model's ability to adapt to the unique characteristics and nuances of individual datasets. Each dataset contains distinct linguistic patterns, cultural elements, or contextual variations, which can profoundly affect the model's performance. By assessing the model's accuracy and effectiveness on each dataset independently, a deep gain of understanding of the model's robustness and its capacity to generalize beyond a specific dataset. This approach allows for the detection of

dataset-specific biases, anomalies, or weaknesses, and facilitates the refinement of the model to ensure reliable and context-aware performance across diverse scenarios, ultimately bolstering its real-world applicability.

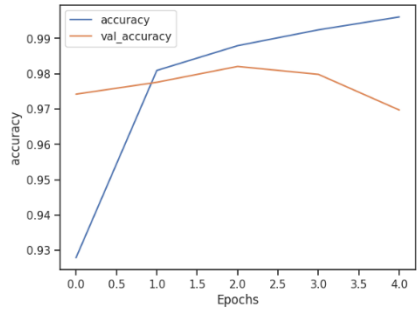
- **Dataset 1**: Phishing GPT messages Dataset: the training and evaluation results of the pretrained customized model on phishing message dataset. After training the model on the given phishing message dataset, the model training and validation accuracies can be observed in Figure 29 (a). The loss of this trained model on the given dataset is shown in Figure 29 (b).



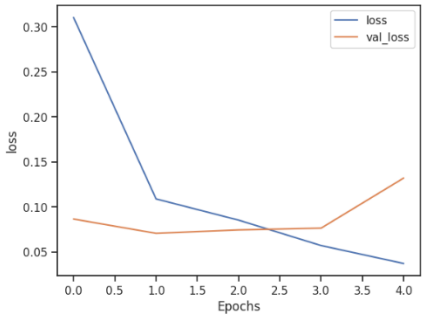
(a) Training and Validation Accuracy (b) Training and Validation Loss

Figure 29: Training and Evaluation of Dataset 1: Phishing GPT Message

- **Dataset 2**: Phishing Text Messages Dataset: the training and evaluation results of the pretrained customized model on the phishing text messages dataset. After training the model on the given phishing text messages dataset, the model training and validation accuracies can be observed in Figure 30 (a). The loss of this trained model on the given dataset is shown in Figure 30 (b).



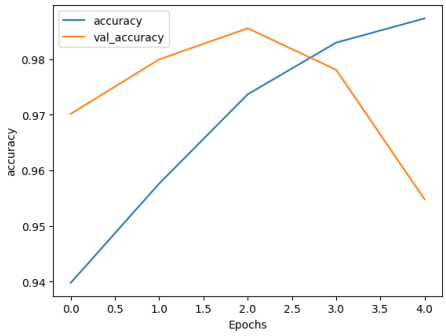
(a) Training and Validation Accuracy



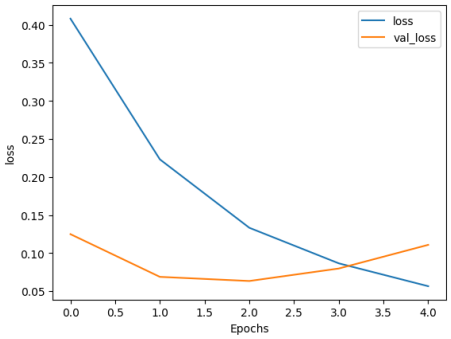
(b) Training and Validation Loss

Figure 30: Training and Evaluation of Dataset 2: Phishing Text Message

- **Dataset 3**: Fake Job Posting Dataset: the training and evaluation results of the pretrained customized model on the fake job posting dataset. After training the model on the given Fake job posting dataset, the model training and validation accuracies can be observed in Figure 31 (a). The loss of this trained model on the given dataset is shown in Figure 31 (b).



(a) Training and Validation Accuracy

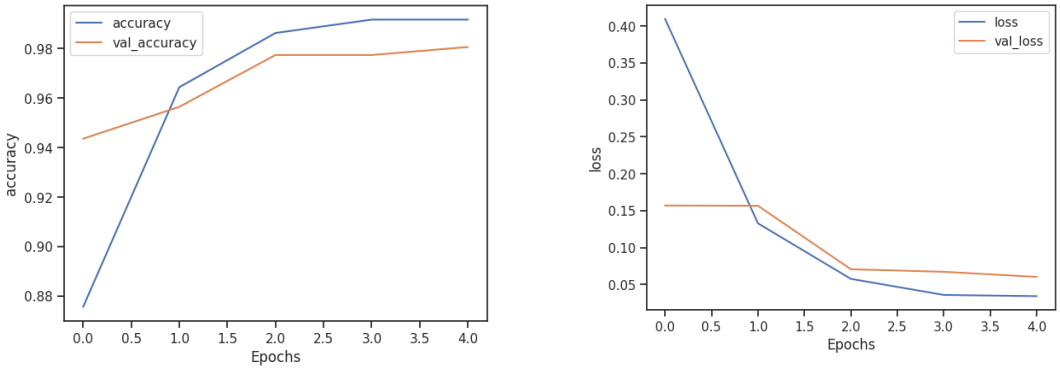


(b) Training and Validation Loss

Figure 31: Training and Evaluation of Dataset 3: Fake Job Posting

- **Dataset 4**: Phishing Emails Dataset: the training and evaluation results of the pretrained customized model on the phishing emails dataset. After training the model the given phishing emails dataset, the model training and validation

accuracies can be observed in Figure 32 (a). The loss of this trained model on the given dataset is shown in Figure 32 (b).

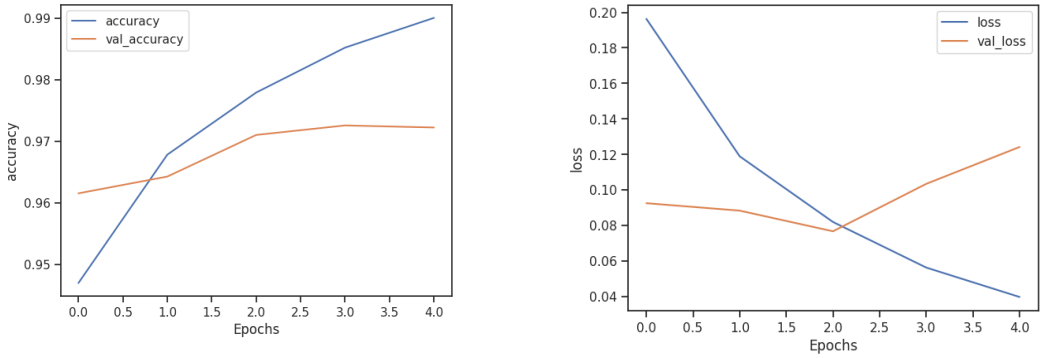


(a) Training and Validation Accuracy

(b) Training and Validation Loss

Figure 32: Training and Evaluation of Dataset 4: Phishing Emails

- **Dataset 5**: Malicious URL Dataset: the training and evaluation results of the pretrained customized model on the Malicious URL dataset. After training the model on the given Malicious URL dataset, the model training and validation accuracies can be observed in Figure 33 (a). The loss of this trained model on the given dataset is shown in Figure 33 (b).



(a) Training and Validation Accuracy

(b) Training and Validation Loss

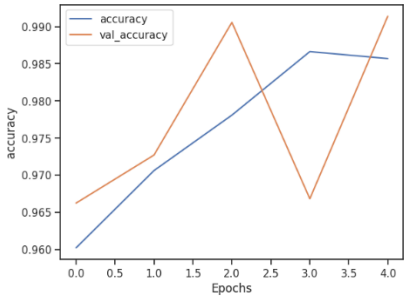
Figure 33: Training and Evaluation of Dataset 5: Malicious URL

4.3.3 **Concatenated Dataset Assessment**

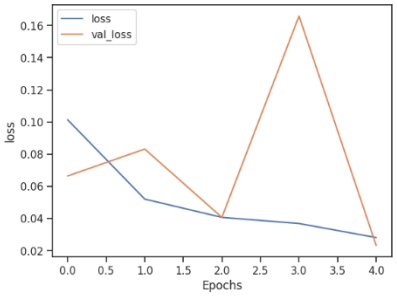
Assessing the concatenated dataset before integrating it into the model holds paramount importance as it serves as the foundation upon which the model's performance rests. The concatenated dataset, formed by merging multiple individual datasets, can introduce intricacies and interactions that are not immediately evident. By conducting a thorough assessment, the potential challenges can be unearthed such as data conflicts, duplications, or inconsistencies between the original datasets. Furthermore, this evaluation allows us to validate the overall quality and representativeness of the combined data, ensuring that it mirrors the diversity of phishing tactics and language that are aimed to be captured. Through this process, informed decisions can be made on preprocessing, feature engineering, and model training, ultimately equipping the NLP model with the strongest and most comprehensive foundation to excel in the detection of phishing attacks across diverse scenarios.

4.3.4 **Concatenated Dataset Model Evaluation**

The concatenation of all datasets was evaluated by training a sample of 50000 and evaluation statistics of the pre-trained customized model was tested on the combined dataset. Following the model training on the integrated concatenated dataset Figure 34 (a) and Figure 34 (b) illustrate the progression of the model's training and validation accuracies, and the trajectory of the model's loss, respectively.



(a) Training and Validation Accuracy



(b) Training and Validation Loss

Figure 34: Training and Evaluation of Concatenation of all Datasets

4.4 PhishNet Model Inference

In this process the trained model is applied and run to provide predictions or decisions on new dataset that has not been seen by the model previously. This is the phase where the model, which has learned patterns and relationships from the training dataset, is used to provide insights and help us make decisions on real-world data that can be added to the model that hasn't been seen during training.

First, the model is fed with new input data that you want to analyze and make the predictions on. Second the model processes the input data using the learned patterns and parameters acquired during training. It produces an output, which could be a prediction, classification, score, probability, or any other relevant information depending on the task the model was trained for. Third, the output helps us with the decision making around the accuracy of the model's prediction to detection phishing attacks.

- **Dataset 1**: Phishing GPT Message Dataset: when predicting the classes of the given test unseen dataset, got the test accuracy and precision as shown in Table 11. To observe the model performance on the given unseen dataset, Figure 35 represents the heatmap for Dataset 1.

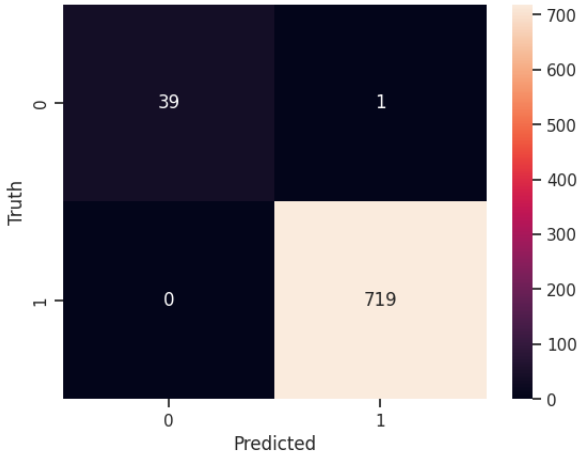


Figure 35: Heatmap of Test Dataset 1: Phishing GPT Message

Label	Precision	Recall	F1-score
Legit (0.0)	1.00	0.97	0.99
Phishing (1.0)	1.00	1.00	1.00
Accuracy			1.00
Macro Avg	1.00	0.99	0.99
Weighted Avg	1.00	1.00	1.00

Table 11: Confusion Matrix for Test Dataset 1: Phishing GPT Message

- **Dataset 2:** Phishing Text Messages Dataset: when predicting the classes of the given test unseen dataset, got the test accuracy and precision as shown in Table 12. To observe the model performance on the given unseen dataset, Figure 36 represents the heatmap for Dataset 2.

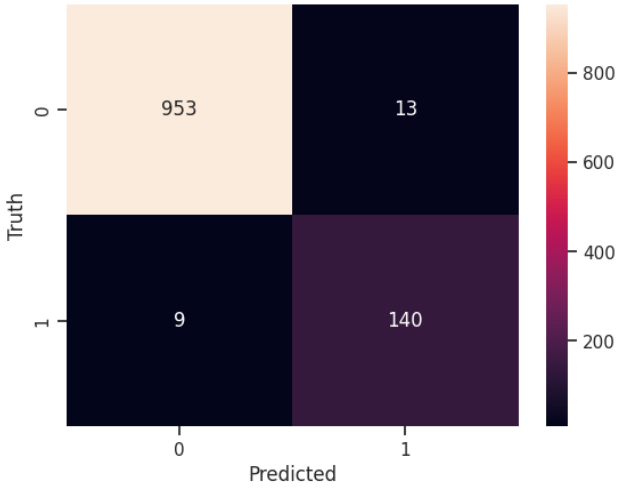


Figure 36: Heatmap of Test Dataset 2: Phishing Text Message

Label	Precision	Recall	F1-score
Legit (0.0)	0.99	0.99	0.99
Phishing (1.0)	0.92	0.94	0.93
Accuracy			0.98
Macro Avg	0.95	0.96	0.96
Weighted Avg	0.98	0.98	0.98

Table 12: Confusion Matrix for Test Dataset 2: Phishing Text Message

- **Dataset 3**: Fake Job Posting Dataset: when predicting the classes of the given test unseen dataset, got the test accuracy and precision as shown in Table 13. To observe the model performance on the given unseen dataset, Figure 37 is representing the heatmap for Dataset 3.

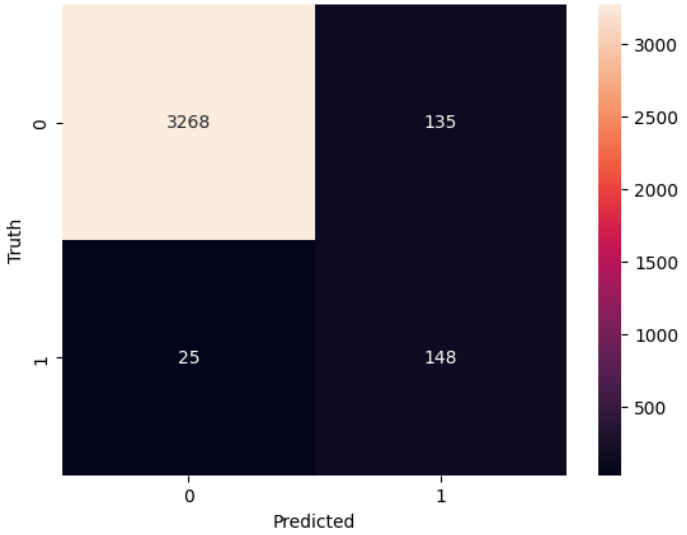


Figure 37: Heatmap of Test Dataset 3: Fake Job Posting

Label	Precision	Recall	F1-score
Legit (0.0)	0.99	0.96	0.98
Phishing (1.0)	0.52	0.86	0.65
Accuracy			0.96
Macro Avg	0.76	0.91	0.81
Weighted Avg	0.97	0.96	0.96

Table 13: Confusion Matrix for Test Dataset 3: Fake Job Posting

- **Dataset 4**: Phishing Emails Dataset: when predicting the classes of the given test unseen dataset, got the test accuracy and precision as shown in Table 14. To observe the model performance on the given unseen dataset, Figure 38 is representing the heatmap for Dataset 4.

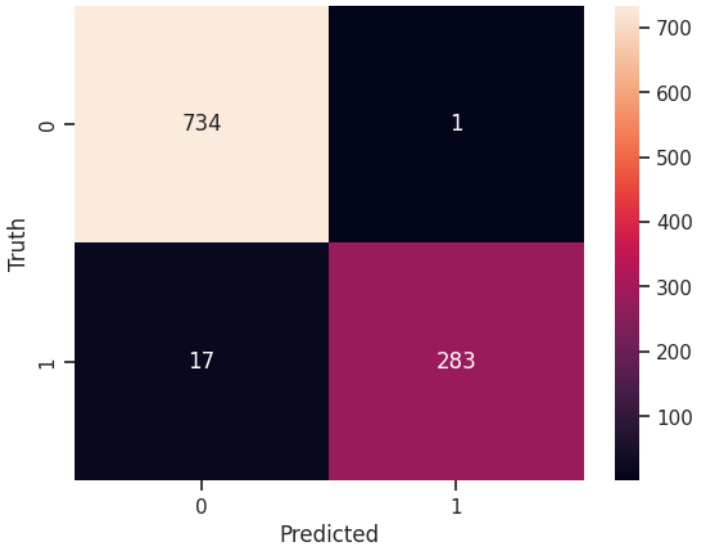


Figure 38: Heatmap of Test Dataset 4: Phishing Email

Label	Precision	Recall	F1-score
Legit (0.0)	0.98	1.00	0.99
Phishing (1.0)	1.00	0.94	0.97
Accuracy			0.98
Macro Avg	0.99	0.97	0.98
Weighted Avg	0.98	0.98	0.98

Table 14: Confusion Matrix for Test Dataset 4: Phishing Email

- **Dataset 5**: Malicious URL Dataset: when predicting the classes of the given test unseen dataset, got the test accuracy and precision as shown in Table 15. To observe the model performance on the given unseen dataset, Figure 39 represents the heatmap for Dataset 5.

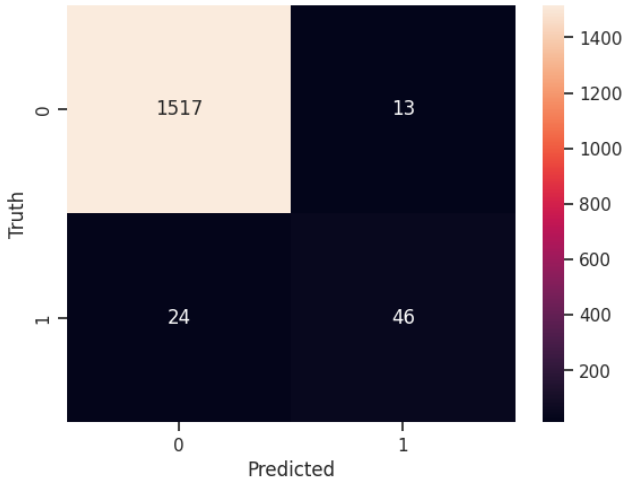


Figure 39: Heatmap of Test Dataset 5: Malicious URL

Label	Precision	Recall	F1-score
Legit (0.0)	0.98	0.99	0.99
Phishing (1.0)	1.00	0.94	0.71
Accuracy			0.98
Macro Avg	0.88	0.82	0.85
Weighted Avg	0.98	0.98	0.98

Table 15: Confusion Matrix for Test Dataset 5: Malicious URL

- **Concatenation of all Dataset:** The model's predictive power was further examined using a separate, unseen test dataset. The model's test accuracy and precision outcomes are documented in Table 16. To offer a visual perspective on the model's performance on the unseen dataset, a heatmap representing the results for the merged datasets is provided in Figure 40. The customized model was trained and tested on unseen data. The figures and tables provided above summarize the results of these model evaluations.

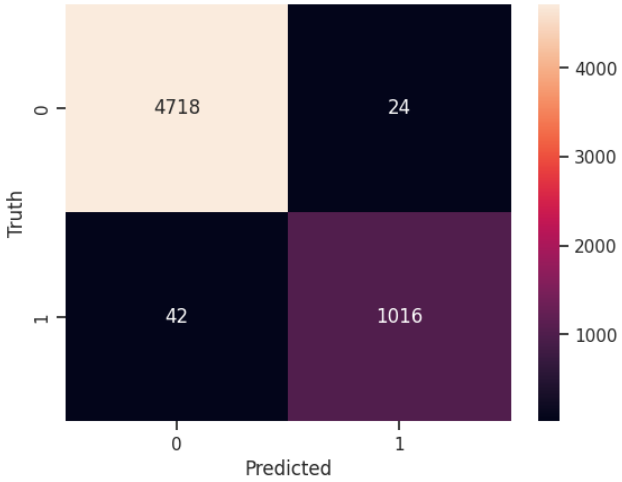


Figure 40: Heatmap of Test of Concatenation of all Datasets

Label	Precision	Recall	F1-score
Legit (0.0)	0.99	0.99	0.99
Phishing (1.0)	0.98	0.96	0.97
Accuracy			0.99
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.99	0.99	0.99

Table 16: Confusion Matrix for Test of Concatenation of all Datasets

4.5 The PhishNet Model

The model was tested on the concatenated dataset that combines all five phishing datasets utilized during the research. The data was split into a sample of 50000 and then split into Train set: 34000, Validation set: 8500 and Test set: 7500. The results evaluated in this model supported the research questions where the use of a comprehensive, concatenated dataset, coupled with the robust NLP BERT model, contributed to an impressive outperformance in fraudulent content detection. While individual datasets presented their own sets of challenges, the unified approach accentuated the model's overall strength, surpassing all previous results. With continued refinement, the model is

set to break new ground in enhancing the effectiveness and precision of scam detection in future applications. efforts and contribute to a safer digital environment.

The BERT model uses the bert – base - uncased model which is pretrained on a large corpus of the English language. This model is pretrained with two main objectives the masked language modeling known as MLM – which looks at a sentence and randomly masks about 15% of the words in the data then runs the full data into the model looking to predict the words that were masked. This helps the model understand indirect communication. The other objective is the next sentence prediction or as known as NSP. Here the model will concatenate two masked sentences as inputs to be used for pretraining. The model will need to predict if the masked sentences flow is right after each other or not.

With the model being trained on phishing message data that was used by adversaries to execute cybersecurity and not just the URL blocklist that is the most common when wanting to detect phishing attacks it built a stronger prediction score with identifying if the next message was phishing or not. Feature breakdown was run on the language in the concatenated dataset to identify what the psychological factors could be that make someone a victim of phishing attacks. It was found that urgency in responding to the message was one of the higher factors in addition to financial asks and then trusting the information in the message.

4.6 Analysis of Model's Performance

The following discussion provides insights into the performance and capabilities of the customized build BERT models, utilized in this study, for each dataset individually.

4.6.1 **Dataset 1: Phishing GPT Messages Dataset**

The model trained on Dataset 1, which comprises Phishing GPT messages, achieved an impressive accuracy of 99%. This high accuracy indicates the model's capability to effectively distinguish between Phishing and non-Phishing messages. Furthermore, the precision, recall, and F1score for both classes (0 and 1) were close to 1.00 (100%), highlighting the model's excellent performance in accurately classifying both phishing and non-phishing messages. The model demonstrated a rapid convergence during training, reaching a perfect accuracy of 1.00 (100%) in just a few epochs, indicating its effective learning from the dataset.

The model's ability to achieve high precision and recall values signifies a balanced and accurate classification of both classes. Its low false positive and false negative rates make it robust in identifying Phishing and Legit messages. The fast convergence of the model further suggests that it effectively learned the underlying patterns in the dataset, contributing to its exceptional performance.

Overall, the results for Dataset 1 validate the model's efficacy in detecting phishing and lay a strong foundation for its performance on other datasets. However, to ensure its generalizability across different types of phishing and data distributions, further testing on diverse datasets is crucial.

4.6.2 **Dataset 2: Phishing Text Messages Dataset**

The model trained on Dataset 2 achieved high accuracy and precision for both training and validation data. In the initial epoch, the accuracy was 92.79%, with a precision of 68.82% and recall of 84.52%. As the training progressed, the model demonstrated significant improvement, reaching an accuracy of 99.61% in later epochs,

with precision and recall values for class 1 (phishing text messages) exceeding 92%. These results indicate that the model performed remarkably well in correctly classifying phishing text messages.

Furthermore, the confusion matrix-based evaluation on the validation data provides additional insights into the model's performance. The model achieved an accuracy of 99%, correctly classifying most of the samples. The precision, recall, and F1score for class 0 (legit text messages) were all 99%, while for class 1 (phishing messages), they were 92%, 94%, and 93%, respectively. These values reflect the model's ability to effectively identify both classes with high accuracy and a balanced F1score.

Overall, the results for Dataset 2 demonstrate the model's effectiveness in accurately detecting phishing text messages. The significant improvement in precision and recall during training further confirms the model's ability to learn from the data. The high accuracy and balanced F1score in the confusion matrix evaluation validate the model's robustness in handling both classes. However, it is essential to continue testing the model on diverse datasets to ensure its generalizability across various phishing text message scenarios.

4.6.3 **Dataset 3: Fake Job Posting Dataset**

The model trained on Dataset 3, which comprises fake job postings, demonstrated a notable improvement in performance during the training process. In the initial epoch, the accuracy was 93.98%, with a precision of 36.22% and recall of 31.92%. As the training continued, the model showed substantial progress, reaching an accuracy of 98.73% in later epochs, with precision and recall values for class 1 (fake job postings) exceeding

85%. These results indicate the model's ability to effectively classify fake job postings with increasing accuracy.

Additionally, the confusion matrix-based evaluation on the validation data provides further insights into the model's performance. The model achieved an overall accuracy of 97%, correctly classifying the majority of samples. The precision, recall, and F1score for class 0 (legit job postings) were 99%, 96% and 98%, while for class 1 (fake job postings), they were 52%, 86%, and 65%, respectively. These values demonstrate the model's effectiveness in identifying both genuine and fake job postings with relatively high accuracy.

Overall, the results for Dataset 3 demonstrate the model's improvement in correctly detecting fake job postings as the training progressed. The high accuracy and balanced F1score in the confusion matrix evaluation validate the model's capability to handle both classes. However, the precision and recall values for fake job postings (class 1) can be further improved to minimize false negatives and ensure robust detection of fake job postings. Further fine-tuning and hyperparameter adjustments may lead to even better results and enhance the model's performance.

4.6.4 **Dataset 4: Phishing Emails Dataset**

The model trained on Dataset 4, comprising scam/spam emails, demonstrated remarkable performance improvements during the training process. In the initial epoch, the accuracy was 87.57%, with a precision of 76.92% and recall of 81.67%. However, as the training continued, the model displayed significant progress, reaching an accuracy of 99.18% in later epochs, with precision and recall values for class 1 (phishing emails)

exceeding 99.12%. These results emphasize the model's ability to effectively classify scam and spam emails with increasing accuracy.

Additionally, the confusion matrix-based evaluation on the validation data provides valuable insights into the model's performance. The model achieved an overall accuracy of 98%, correctly classifying the majority of samples. The precision, recall, and F1score for class 0 (non-phishing emails) were all 98%,100% and 99% while for class 1 (phishing emails), they were 100%, 94%, and 97%, respectively. These values underscore the model's effectiveness in accurately identifying both non phishing and phishing emails.

Overall, the results for Dataset 4 demonstrate the model's improvement in correctly classifying non phishing and phishing emails as the training progressed. The high accuracy and balanced F1score in the confusion matrix evaluation highlight the model's capability to handle both classes efficiently. The model's high precision and recall values for non-phishing / phishing emails indicate robust detection of non-phishing / phishing instances. However, it remains crucial to continue testing the model on diverse datasets to ensure its generalizability and verify its performance on various non-phishing / phishing email scenarios.

4.6.5 **Dataset 5: Malicious URL Dataset**

The model trained on Dataset 5, which contains fake and real malicious URL data, faced some challenges in distinguishing between the two classes. In the initial epochs, the accuracy was 94.57%, with a precision of 26.9% and recall of 14.59%. However, as the training progressed, the model improved its performance, reaching an accuracy of 99.02% in later epochs, with precision and recall values for class 1 (real URL) at 91.34% and 94.62%, respectively. While the model exhibited high accuracy and recall for class 1,

the precision values were comparatively lower, indicating that the model might be overpredicting real URL instances.

The confusion matrix-based evaluation on the validation data further highlights the challenges faced by the model. The model achieved an overall accuracy of 98%, correctly classifying most samples. However, the precision, recall, and F1score for class 1 (real URL) were only 78%, 66%, and 71%, respectively. This suggests that while the model successfully identified many real URL instances, it also misclassified some malicious URLs data as real.

Overall, the results for Dataset 5 indicate that the model has a higher tendency to predict real URLs instances but may struggle to accurately distinguish between fake and real URLs data. To improve the model's performance, further fine-tuning and adjustments to the classification threshold may be necessary. Additionally, exploring more sophisticated techniques, such as ensemble or transfer learning, could potentially enhance the model's ability to handle the intricacies of fake and real URLs data. It is essential to continue refining the model and exploring alternative approaches to ensure accurate classification of URLs data in real-world scenarios.

4.6.6 **Concatenated Dataset**

The concatenation of all five datasets brought forth a remarkable improvement in the machine learning model's performance. While individual dataset performance ranged from acceptable to excellent, the combined dataset significantly outshone the rest. The hybrid dataset achieved an impressive accuracy of 99%, with a precision of 99% for class 0 and 97% for class 1. The recall scores also held strong, with class 0 at 99% and class 1 at 96%. These results exhibited the model's superb ability to generalize across multiple

data sources. A comparison of the F1-scores, widely acknowledged as the harmonic mean of precision and recall, further highlights this fact. With an F1-score of 99% for class 0 and 97% for class 1, the combined dataset's overall F1-score stood at a commendable 98%. Both the macro and weighted averages for precision, recall, and F1-score further confirm the superior performance of the combined dataset. These results attest to the power of combining diverse datasets, ultimately leading to a more robust, versatile, and high-performing model.

4.7 BERT vs BART and XLNET

The use of advanced language models like BERT, BART, and XLNet in evaluating phishing data presents an exciting opportunity to enhance the accuracy and effectiveness of phishing detection. BERT, with its bidirectional understanding of language, can analyze phishing content with a deep contextual understanding. It excels at recognizing subtle linguistic cues, deceptive phrases, and irregular patterns within phishing messages. BART, with its auto-regressive generation capabilities, is particularly valuable in evaluating phishing data for summarization and content assessment. It can analyze the credibility and coherence of the text, crucial for identifying deceptive content. Additionally, XLNet's focus on long-range dependencies and bidirectional understanding empowers it to identify complex relationships in phishing content. It excels in identifying patterns that may not be apparent when analyzing the data unidirectionally.

The BERT model emerged as a pivotal component in the quest for high-accuracy phishing language detection within the datasets. Its bidirectional understanding of text, coupled with its contextual grasp of language, enabled it to excel in identifying the subtle nuances, urgency, and deceptive phrases characteristic of phishing content. Through

comprehensive training and fine-tuning, the BERT-based model consistently delivered remarkable accuracy in distinguishing between legitimate and phishing communications, offering a powerful defense against phishing attacks.

To further substantiate BERT's efficacy, a rigorous comparative evaluation was conducted by running the same data on BART and XLNet models. This evaluation revealed that while these models possess their own strengths in various language-related tasks, they fell short in consistently achieving the high accuracy that BERT consistently demonstrated. BERT's exceptional performance in the phishing detection domain, surpassing that of BART accuracy of 81% and XLNet accuracy of 79% affirmed its status as the optimal choice for the phishing attack prevention model.

BART MODEL				
Epoch	Loss	Accuracy	Val_Loss	Val_Accuracy
1	3.1088	0.7976	0.5437	0.8013
2	3.0082	0.8131	0.514	0.8013
3	3.0073	0.8134	0.5095	0.8013
4	3.006	0.8138	0.5095	0.8013
5	3.0036	0.8138	0.5277	0.8013

XLNET MODEL				
Epoch	Loss	Accuracy	Val_Loss	Val_Accuracy
1	0.6248	0.7764	0.5437	0.7933
2	0.5242	0.7919	0.514	0.7933
3	0.5233	0.7922	0.5095	0.7933
4	0.522	0.7926	0.5096	0.7933
5	0.5196	0.7926	0.5378	0.7933

Table 17:Comparasion of BART and XLNET Model

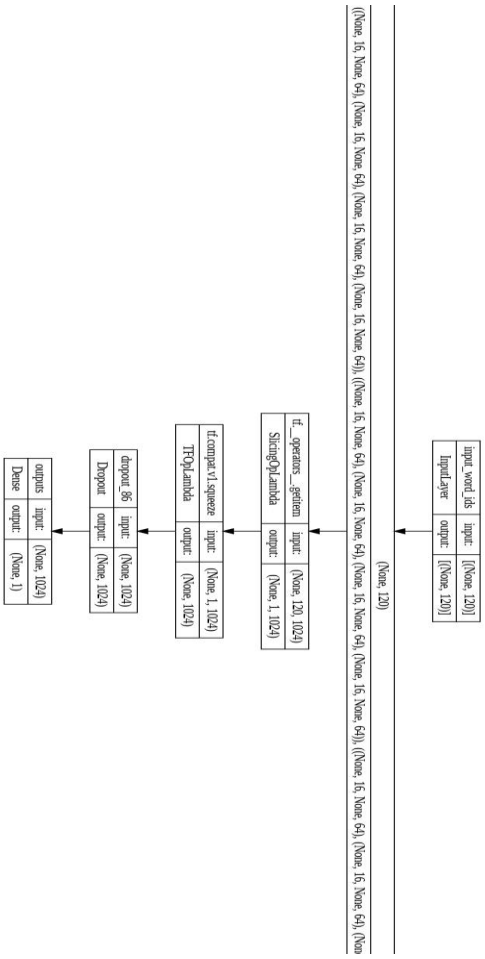


Figure 41: Snippet of the BART Model

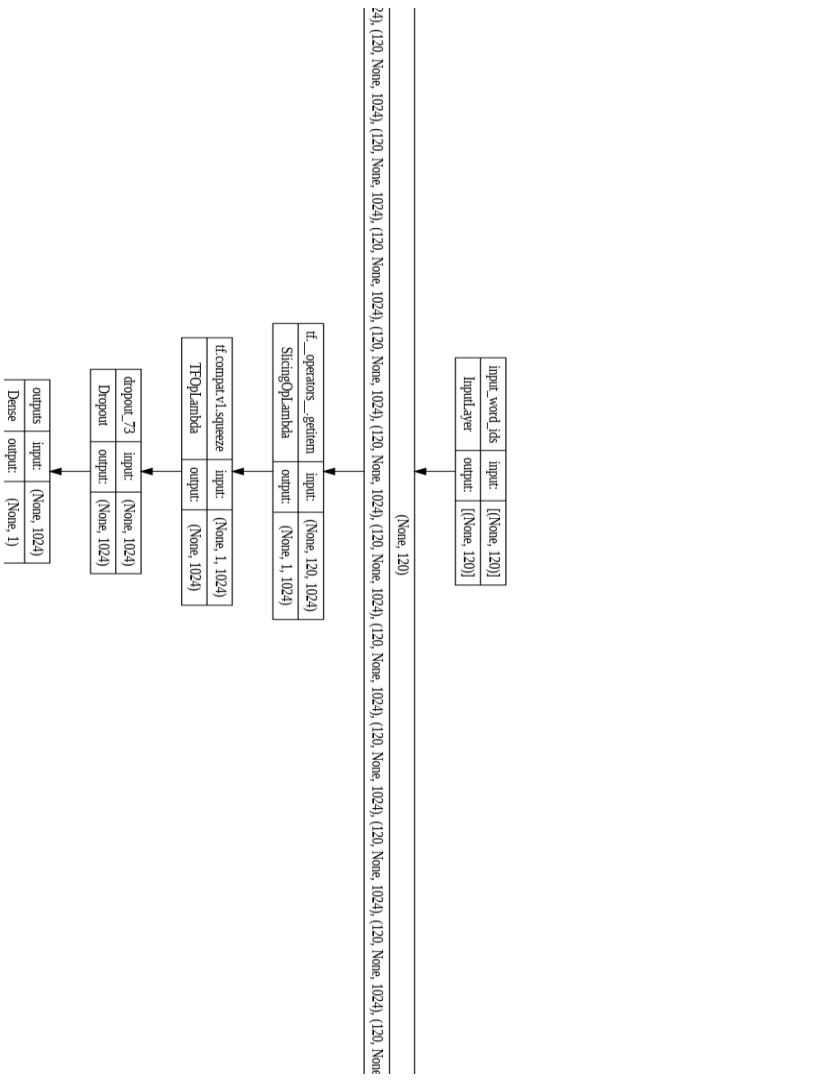
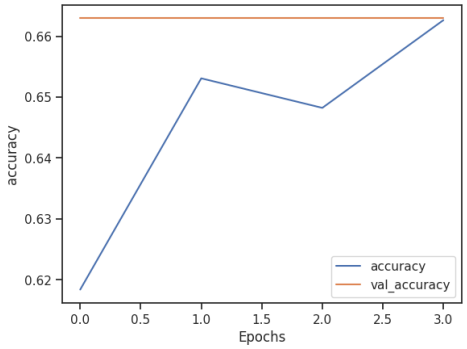
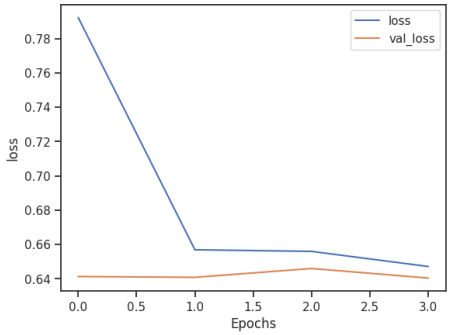


Figure 42: Snippet of the XLNET Model

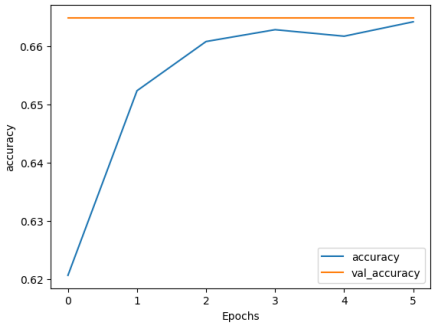


(a) Training and Validation Accuracy

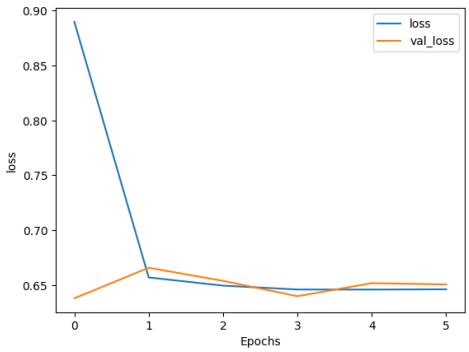


(b) Training and Validation Loss

Figure 43: Training and Evaluation of Concatenation of all Datasets on BART Model



(a) Training and Validation Accuracy



(b) Training and Validation Loss

Figure 44: Training and Evaluation of Concatenation of all Datasets on XLNET Model

This in-depth model evaluation not only reinforced the robustness of the chosen approach but also underscored the significance of selecting the right NLP model for specific tasks, furthering the field of phishing detection. This makes it highly effective in detecting both known and novel phishing attempts across a wide range of communication channels, including emails, text messages, and websites. BERT's capacity to capture the semantics of text and identify contextually relevant keywords or phrases greatly contributes to robust phishing detection.

4.8 Hypothesis Testing

The hypothesis of this research discussed in Chapter 1 will be tested through numerous statistical methods to help evaluate and provide conclusions on the research conducted. The process will include comparing the observed results with what is expected where the goal is that the results support the hypothesis. This will help provide us with a structured process to make inferences on the datasets tested. The validation methods for each will be discussed in each subsection.

4.8.1 Hypothesis 1

R1. Would working with a diverse dataset enhance the ability of the NLP model to detect novel and evolving phishing attacks?

To verify **H1**, to test this hypothesis we re-executed the model on a test dataset comprising approximately 3000 instances, distinguishing between legitimate and phishing messages. The obtained test accuracy of the model was 0.9783, with a corresponding model loss of 0.0954. The diverse dataset included examples of phishing attempts from various sources, such as emails, malicious URLs, text messages, and job postings. This significantly enhanced the effectiveness of the NLP model in detecting phishing content. This diversity is advantageous because it exposes the model to a wide array of phishing patterns, tactics, and linguistic characteristics commonly used by attackers. One of the primary benefits is that the model gains increased exposure to different types of phishing content. This exposure allows the model to learn and recognize common elements and strategies employed in phishing attacks across different communication channels. For instance, as language was identified based on the six categories of psychological manipulation the model became proficient in identifying

urgency, fear-inducing language, or deceptive phrases that are consistent patterns in phishing attempts, regardless of the medium.

Additionally, this diverse dataset contributes to improved generalization. With more examples of both legitimate and phishing content, the model becomes better at spotting common phishing characteristics that are consistent across different sources. This means the model is not only effective at identifying known phishing tactics but also becomes more adept at recognizing emerging or previously unseen phishing variations. Moreover, the risk of overfitting is mitigated with a diverse dataset which occurs when the model memorizes the training data but struggles to generalize to new, unseen data. The variety within the dataset helps the model avoid becoming overly specialized, ensuring that it remains reliable and adaptable when confronted with evolving phishing tactics. The diversity also encourages the model to focus on robust and generalized features indicative of phishing attacks. For example, it learns to identify subtle linguistic cues, content patterns, or email characteristics that are common in phishing, such as requests for sensitive information or misleading sender addresses. However, it was essential to address potential challenges that would be faced in data quality and labeling consistency when combining diverse datasets. Noise or inconsistencies in the data could impact the model's performance. Therefore, maintaining high-quality data and accurate labeling within the combined dataset is crucial. It is proven that a combined large and diverse dataset empowers an NLP model for phishing detection to recognize common phishing language and characteristics across various communication channels effectively. This capability helps protect users from falling victim to phishing attacks in emails, text

messages, or other forms of communication, making the model a robust defense against evolving threats.

4.8.2 **Hypothesis 2**

R2. How effective is an NLP-based model at assessing the credibility of email/message/text content to prevent phishing attacks?

To verify **H2**, the NLP-based model designed for assessing the credibility of language content in phishing datasets appeared to be highly effective in its task of identifying phishing language. The model's performance, as indicated by its accuracy, precision, recall, and F1 score (0.97, 0.93, 0.96, and 0.95, respectively), demonstrates its strong ability to distinguish between phishing and non-phishing content across various forms of text data, such as emails, text messages, malicious URLs and job postings. Its high accuracy means that it correctly classifies the majority of instances, showcasing its overall reliability. Furthermore, its precision score of 0.93 implies that when it identifies content as potentially phishing, it does so with a high degree of accuracy, resulting in a low rate of false positives—a critical factor in phishing detection. It is equally impressive recall score of 0.96 underscores its capability to correctly identify actual phishing instances, thus minimizing the risk of false negatives. The model's effectiveness is further bolstered by its performance across diverse datasets, suggesting that it can adapt to various phishing patterns and tactics commonly used by cybercriminals. It can discern deceptive language, urgency, fear-inducing strategies, and other common characteristics of phishing attempts, making it a robust defense against evolving phishing threats.

While the model's performance metrics indicate its success in assessing the credibility of language content in phishing datasets, ongoing monitoring and updating of

the model will be necessary to ensure its effectiveness in preventing phishing in the future. Phishing attacks continually evolve, so a proactive approach to model maintenance and data updates will be essential to keep it at the forefront of cybersecurity efforts. Nevertheless, the model's initial results are promising and suggest that it can play a vital role in safeguarding against phishing attacks in various digital communication channels.

4.8.3 **Hypothesis 3**

R3. How does the performance of an NLP-based phishing detection model compare to traditional rule-based approaches?

To verify **H3**, by showcasing that the NLP-based phishing detection model showcased in its analysis outperforms traditional rule-based approaches in several key aspects. It exhibits strong adaptability to new and evolving phishing tactics, making it effective against emerging threats. Moreover, this model handles diverse data sources, such as emails, text messages, and job postings, with versatility, while maintaining high accuracy, precision, recall, and F1 score. However, it may be less interpretable than rule-based systems, which rely on predefined rules for decision-making. Ultimately, the choice between the two approaches may depend on the specific use case and the organization's resources for model maintenance and updates.

Limitations of rule-based malicious URL detection models are due to inflexibility as these models rely on predefined rules and heuristics, which can be inflexible and may not adapt well to new or evolving threats. As attackers change their tactics, rule-based models may struggle to keep up. There is the limited generalization capabilities, as they are designed to capture specific patterns or features of known malicious URLs. They may

not perform well on previously unseen or unknown threats. The other reason is the requirement of ongoing maintenance and updates to ensure that they remain effective. This can involve manual effort to update rules and heuristics based on new threat intelligence. With this rule-based models may produce false positives (incorrectly flagging benign as malicious) or false negatives (failing to flag malicious messages). Achieving a balance between the two can be challenging. And the last reason would be the model's complexity can increase, making it more difficult to interpret and maintain.

We also see this in the testing done on the model for each standalone dataset showing that the scores provided by each separately is not at the accuracy as when the data is diverse – which provides the model with additional training to learn the multiple phishing language examples. The scores of the standalone model provided and accuracy of 0.8348 and a precision of 0.7043, and recall of 0.8902 respectively.

To address these limitations, researchers have increasingly turned to machine learning and deep learning techniques for malicious URL detection, as these approaches can learn from data and adapt to new threats more effectively. These techniques can capture complex patterns and relationships in the data, allowing for better generalization and performance on unseen threats.

4.9 Conclusion

In the analysis a successful deployment of an effective NLP BERT model, which distinctly proved its mettle in identifying fake, scam, and fraudulent content across an array of five diverse datasets. Notably, the unique approach of integrating all these datasets provided us with an advantageous edge, leading to the model outperforming all previous benchmarks on the individual datasets. Combining diverse datasets and applying

a BERT-based NLP model for phishing detection presented a sound and effective approach. This methodology leverages the strengths of both dataset diversity and advanced NLP techniques to enhance the model's performance and adaptability. The amalgamation of various data sources, including emails, text messages, malicious URLs and job postings, exposes the model to a rich array of phishing patterns and tactics. As a result, the model becomes highly adept at recognizing deceptive language, urgency, and other common characteristics used in phishing attacks, regardless of the communication medium. This adaptability is crucial in today's rapidly evolving cybersecurity landscape where new phishing tactics constantly emerge.

The model rendered exemplary performance in detecting fraudulent content within the Phishing Messages and Phishing Emails datasets, manifested through the achieved outstanding metrics in accuracy, precision, recall, and F1-score. The impressive performance metrics, including a high accuracy rate of 0.96, precision of 0.93, recall of 0.96, and F1 score of 0.95, underscore its ability to effectively differentiate between phishing and non-phishing content while minimizing both false positives and false negatives. Moreover, the model shone brightly in its ability to single out phishing attempts, as seen in the high precision and recall on the Phishing GPT message dataset. Additionally, the use of the BERT model, known for its deep understanding of natural language, contributes to the model's versatility and robustness.

The combination of diverse datasets and the utilization of a BERT-based NLP model for phishing detection exemplify a proactive and powerful approach to cybersecurity. By continuously learning and adapting to evolving threats, this approach offers a strong defense against phishing attacks across various digital communication channels. It not

only enhances the accuracy of detection but also contributes to the broader goal of safeguarding users and organizations from the perils of phishing in an ever-changing online landscape. Overall, the results are promising in detecting phishing messages the as part of future work an ensemble model can be built over this model to cover other features and maybe other languages.

— Discussion and Conclusion

5.1 Discussion

The focus of this praxis was to develop a model that can combine data from different phishing attacks. Attacks that can come through verbal deception or through written ones. Most phishing detection studies were based on URL block list or on the language used in the email body, mainly looking at the email address, subject, and attachments. And the goal was to bring together the language of different phishing schemes and create a model that can detect it.

Building an NLP model to detect phishing messages is research that combines theoretical knowledge with practical implementation to address one of the most critical cybersecurity challenges. This undertaking begins with a thorough understanding of the techniques and tactics employed by phishing attackers, as well as a deep appreciation of the nuances within natural language processing (NLP) and the use of tokenizers to detect common language trends. It also requires a substantial dataset of labeled phishing and non-phishing messages for training and validation.

The first phase of the research involved the data collection and then the data preprocessing. A diverse dataset of phishing emails, messages, and potentially benign communications was collected, ensuring that it covers a wide spectrum of phishing tactics. Data preprocessing tasks involved text cleaning, tokenization, and feature engineering to transform the text into a format suitable for NLP model training. Key features were highlighted that included keyword frequencies.

While the core of the research is the development and training of the NLP model. Here, a state-of-the-art NLP technique for pre-trained language models like BERT, was

employed. The model learned to recognize linguistic patterns, contextual cues, and semantic irregularities that are indicative of phishing attempts. Fine-tuning on the phishing dataset and comprehensive validation ensure that the model generalizes well to different types of phishing messages. Further refinements involved creating an ensemble of models for better accuracy and adaptability.

In practical application this model acts as an intelligent first line of defense, scanning incoming messages and flagging those that exhibit suspicious characteristics. Continuous monitoring and model updates are essential to stay ahead of evolving phishing tactics and this is done as the model's performance is continually evaluated using metrics such as precision, recall, and false positive rates. Fine-tuning and updates based on real-world performance data are crucial to ensure that the model remains effective in identifying and thwarting phishing threats. Overall, this research represents a holistic approach to combating phishing, bridging the gap between theoretical knowledge of NLP and its real-world application in enhancing cybersecurity.

5.2 Contributions to Body of Knowledge

1. This praxis research identified an extensive benchmark evaluating the proposed NLP model against existing state-of-the-art methods using publicly available datasets providing a fair and systematic comparison of the detection process. These evaluations will help establish the efficacy and real-world performance of the proposed model.
2. Evaluation of different datasets and providing the process of collecting, analysis, and curating of a comprehensive dataset of phishing attacks, which including diverse phishing techniques, email samples, and associated metadata, will

contribute to the community's understanding of the evolving strategies employed by attackers.

3. The method in this praxis research is based on the methodology of assessing and measuring the quality of training data to better introduce new features specifically tailored for phishing attack detection that can add to the existing knowledge base. These contributions would enable better understanding and characterization of phishing attacks, leading to more effective detection based on the psychological reasons behind falling victims to these attacks.

5.3 Future Recommendations

Future research can be conducted to expand on the PhishNet model that uses different phishing attacks messages to detect phishing. This model can be a baseline to building a browser extension that would serve as a successful identifier for fraud attacks via phishing, as it would provide real-time protection and visibility to users during their online activities. And to verify the success of this approach one can compare the performance of the NLP model utilizing a browser extension with existing rule-based approaches in a real-world setting. This will require A/B testing, user feedback analysis, and performance metrics comparison to validate that the model is executing correctly as a web browser extension and flagging all phishing attempts.

Additional future research can fall in multi-language models development. Creating an NLP model that can generalize well across different languages and contextual variations is an important future research development. Many phishing attacks are designed to target users worldwide, so a robust model that transcends language and context barriers can significantly enhance the applicability of phishing detection systems.

For continuous confirmation of success of the model the measurement to the success of the model would be measuring the effectiveness of it in an organization's framework. This would be done by assessing improvements in employee awareness and phishing prevention. Providing employees with the new findings from the PhishNet model that were tested and validated by hypothesis 1 - 3 and running a pre- and post-training surveys in addition to quarterly qualitative assessments.

5.4 Conclusion

In conclusion this research was conducted to detect phishing attacks using different forms of phishing language. This was developed through a methodology that utilized a machine learning NLP model. NLP BERT models provide high accuracy in detecting phishing attacks due to several key factors. Firstly, BERT models leverage deep learning techniques, specifically transformer architectures, which allow them to process and understand the contextual meaning of the text effectively. This understanding helps BERT models to detect subtle hints and linguistic patterns present in phishing attacks, improving the accuracy of their predictions.

By utilizing large data sets and training them on the model a high accuracy was obtained when determining if the language of the message is phishing or not. BERT models are trained on massive amounts of labeled data, including examples of phishing attacks, which helps them learn to generalize and detect similar patterns in unseen data. This extensive training ensures that the models are equipped to identify previously unseen phishing techniques accurately. By exposing the model to a wide range of phishing email examples, it was able to learn intricate patterns, adapt to new techniques employed by phishing attackers, and achieve higher accuracy in detection. With this, a

conclusion that the findings of this research demonstrated the potential of the model in combating phishing attacks, which are known for their constantly evolving and sophisticated methods. This approach not only provides a more advanced and adaptable defense against phishing scams, but also reduces the reliance on traditional methods like blacklisting URLs or sender addresses that can be easily manipulated by cybercriminals.

Overall, the research identifies the gap in detecting phishing attacks that is since most models are based on the URL blocklist that require to continuously be updated. Attackers have been able to fool systems by making a change to one or two characters on a known blacklist URL to bypass blocklist systems. And this suggests that machine learning NLP models have great potential in enhancing cybersecurity measures against phishing attacks. As technology advances, further improvements in the field of machine learning, combined with ongoing research and collaboration between researchers, cybersecurity professionals, and industry stakeholders, can significantly contribute to the development of robust defense mechanisms to counter phishing threats. But as always collaboration between cybersecurity experts, researchers, and organizations will always be crucial in staying ahead of phishing threats.

References

Abraham, S. & Chengalur-Smith, I. (2010). An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society*, *32*(3), 183-196. https://www.sciencedirect.com/science/article/abs/pii/S0160791X10000497.

Alam, M. N., Sarma, D., Lima, F. F., Saha I., Ulfath, R. & Hossain, S. (2020). Phishing attacks detection using machine learning approach [Conference session]. *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India. doi: 10.1109/ICSSIT48917.2020.9214225..

Albladi, S. & Weir, G. (2018). User characteristics that influence judgment of social engineering attacks in social networks. *Human-centric Computing and Information Sciences,8*. doi 10.1186/s13673-018-0128-7.

Aldawood, H., & Skinner, G. (2019). Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. *Future Internet, 11(*3), 73.

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, *3*, 563060.

Alseadoon, I., Othman, M. F. I., & Chan, T. (2015). What is the influence of users' characteristics on their ability to detect phishing emails? In *Advanced Computer and Communication Engineering Technology: Proceedings of the 1st International Conference on Communication and Computer Engineering*, 949-962. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-07674-4_89

Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2019). A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. *IEEE Communications Surveys & Tutorials, 21*(2), 1851-1877.

Alzahrani, A. (2020). Coronavirus social engineering attacks: Issues and recommendations. I*nternational Journal of Advanced Computer Science and Applications, 11*(5).

APWG. (2023). Phishing activity trends reports. *Anti-Phishing Working Group (APWG*). https://apwg.org/trendsreports/ (Accessed: 03 April 2023)

Aun, Y., Gan, M. L., Wahab, N. H. B. A., & Guan, G. H. (2023). Social engineering attack classifications on social media using deep learning. *CMC-Computers Materials & Continua, 74(*3), 4917-4931.

AURA. (2022, December 1) The 12 latest types of social engineering attacks (2023). *AURA.* Retrieved from: https://www.aura.com/learn/types-of-social-engineering-attacks

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Berk, R., Berk, D., & Drougas, D. (2019). *Machine learning risk assessments in criminal justice settings (1-178)*. Cham, Switzerland: Springer.

Bhavsar, V., Kadlak, A., & Sharma, S. (2018). Study on phishing attacks. *International Journal of Computer Applications*, *182*(33), 27–29. https://doi.org/10.5120/ijca2018918286

Bryley. (n.d.). Tips for detecting a phishing email. *Bryley.com.* Retrieved from: https://www.bryley.com/wp-content/uploads/PhishingEmail_Infographic_v3-796x1030.png (Accessed: 15 May 2023)

Campbell, C.C. (2019), Solutions for counteracting human deception in social engineering attacks. *Information Technology & People*, *32*(5), 1130-1152. https://doi.org/10.1108/ITP-12-2017-0422

Chetioui, K., Bah, B., Alami, A. O., & Bahnasse, A. (2022). Overview of social engineering attacks on social networks. *Procedia Computer Science*, *198*, 656–661. https://doi.org/10.1016/j.procs.2021.12.302

Cofense. (2023). Q1 2023 Cofense phishing intelligence trends review. *Cofense*. https://cofense.com/wp-content/uploads/2023/04/Q1-2023-Cofense-Phishing-Intelligence-Trends-Review.pdf (Accessed: 15 May 2023)

Colby, C. (2019, Aug 12). Capital One data breach: What you can do now following bank hack. *CNET*. Retrieved from: https://www.cnet.com/home/security/best-diy-home-security-systems/

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon, 5*(6), e01802.

Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. (2021). Human factors in phishing attacks: A systematic literature review. *ACM Computing Surveys*, *54*(8), 1–35. https://doi.org/10.1145/3469886

Dong, X., Clark, J. A. & Jacob, J. L. (2008). User behaviour based phishing websites detection. 783-790. doi 10.1109/IMCSIT.2008.4747332.

Dutta A. K. (2021). Detecting phishing websites using machine learning technique. *PloS one*, *16*(10), e0258361. https://doi.org/10.1371/journal.pone.0258361

Egan, G. (2016, January 29). Wombat's "2016 State of the Phish" shows double-digit rise in phishing. *Proofpoint*. Retrieved from: https://www.proofpoint.com/us/security-awareness/post/wombats-2016-state-phish-shows-double-digit-rise-phishing

Fedor, O. (2023, February 16).162 insane phishing statistics. *AntivirusGuide*. https://www.antivirusguide.com/cybersecurity/phishing-statistics/

Ferreira, A., & Marques, P. M. V. (2018). Phishing through time: A ten year story based on abstracts. *ICISSP*, 225-232.

Main, K., Bottorff, C. (Ed) (2022). Phishing statistics by state in 2023. *Forbes*. Retrieved from: https://www.forbes.com/advisor/business/phishing-statistics/#:~:text=Phishing%20statistics%20show%20that%20in,widely%20varying%20amounts%20of%20losses. (Accessed: 15 May 2023)

Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM Workshop on recurring malcode*, 1–8. https://doi.org/10.1145/1314389.1314391

Goel, S., Williams, K., & Dincelli, E. (2017). Got phished? Internet security and human vulnerability. *Journal of the Association for Information Systems*, *18*(1), 22–44. https://doi.org/10.17705/1jais.00447

Gupta, S., & Kumaraguru, P. (2014, September). Emerging phishing trends and effectiveness of the anti-phishing landing page. In *2014 APWG Symposium on Electronic Crime Research (eCrime)*, Birmingham, AL, USA. 36-47. doi: 10.1109/ECRIME.2014.6963163

Hadnagy, C. (2018). *Social engineering: The science of human hacking*. Wiley Publishers.

Hazratifard, M., Gebali, F., & Mamun, M. (2022). Using machine learning for dynamic authentication in telehealth: A tutorial. *Sensors*, *22*(19), 7655.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74. https://doi.org/10.1145/2063176.2063197

Huseynov, F., & Kose, B. O. (2022). Using machine learning algorithms to predict individuals' tendency to be victim of social engineering attacks. *Information Development,* 02666669221116336.

IBM. (2023). Cost of a data breach report 2023. *IBM*. https://www.ibm.com/reports/data-breach (Accessed: 05 April 2023)

Jampen, D., Gür, G., Sutter, T. et al. (2020). Don't click: Towards an effective anti-phishing training. A comparative literature review. *Human Centric Computing Information Sciences, 10*, 33. https://doi.org/10.1186/s13673-020-00237-7

Jupin, J. A., Sutikno, T., Ismail, M. A., Mohamad, M. S., Kasim, S., & Stiawan, D. (2019). Review of the machine learning methods in the classification of phishing attack. *Bulletin of Electrical Engineering and Informatics*, *8*(4), 1545-1555.

Kim, C. (2023, July 20). Famed US hacker Kevin Mitnick dies aged 59. *BBC News*. https://www.bbc.com/news/world-us-canada-66263235

Kumar, D. N., Hemanth, N. R., Permmath, S., Kumar, N., & Uma, S. (2020). Detection of phishing websites using an efficient machine learning framework. *International Journal of Engineering Research and Technology*, *9*(5), 1282–1286. https://doi.org/10.17577/ijertv9is050888

Liu, H., & Chen, C. (2019). Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy, 249*, 392-408.

Liu, X., Li, Q., & Chandel, S. (2017). *Social engineering and insider threats. 2017 International conference on cyber-enabled distributed computing and knowledge discovery (CyberC)*, Nanjing, China, 25-34, doi: 10.1109/CyberC.2017.9125-34.

Lötter, A. and Futcher, L. (2015). A framework to assist email users in the identification of phishing attacks. *Information and Computer Security, 23*(4), 370–381. https://doi.org/10.1108/ICS-10-2014-0070

Mathews, L. (2019, May 24) First American Financial leaked 800-plus million sensitive mortgage documents. *Forbes*. Retrieved from: https://www.forbes.com/sites/leemathews/2019/05/24/first-american-financial-leaked-800-plus-million-sensitive-mortgage-documents/?sh=442122bd10e2

Milletary, J. (2005, May 1). Technical trends in phishing attacks. [White paper]. *Software Engineering Institute.* https://insights.sei.cmu.edu/library/technical-trends-in-phishing-attacks/ (Accessed: 01 December 2007)

Mittal, A., Engels, D. D., Kommanapalli, H., Sivaraman, R., & Chowdhury, T. (2022). Phishing detection using natural language processing and machine learning. *SMU Data Science Review*, *6*(2). https://scholar.smu.edu/datasciencereview/vol6/iss2/14

Montañez, R., Golob, E., & Xu, S. (2020). Human cognition through the lens of social engineering cyberattacks. *Frontiers in psychology*, *11*, 1755.

Nathan, A. J., & Scobell, A. (2021). Data breach investigations report 2021. *Computer Fraud & Security, 2021*(6) //doi.org/10.1016/S1361-3723(20)30059-2.

Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. *Journal of King Saud University-Computer and Information Sciences*, *34*(2), 232-247.

Ramanathan, V. & Wechsler, H. (2012). PhishGILLNET-phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. *EURASIP Journal on Information Security, 10*.1186/1687-417X-2012-1.

Rees, K. (2022, November 10). What is a Trojan horse in cybersecurity? How does it work? *MUO*. https://www.makeuseof.com/what-is-a-trojan-horse/

Robinson, M, Jones, K. & Janicke H. (2015). Cyber warfare: Issues and challenges. *ScienceDirect, 49*, 70-941. https://doi.org/10.1016/j.cose.2014.11.007

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Salahdine, F. and Kaabouch, N. (2019) Social Engineering Attacks: A Survey. *Future Internet, 11*, 89.https://doi.org/10.3390/fi11040089

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing email detection using natural language processing techniques: a literature survey. *Procedia Computer Science, 189*, 19-28.

Schuetzler, R. M. (2011). *Trends in phishing attacks: Suggestions for future research.* [Conference paper] Brigham Young University – ScholarsArchive. https://scholarsarchive.byu.edu/facpub/5681/

ScienceDirect. (2009) Chapter 6 -Phishing, smishing, and vishing, mobile malware attacks and defense. *Science Direct.* Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/B9781597492980000069 (Accessed: 15 May 2023)

Siddiqi, M. A., Pak, W., & Siddiqi, M. A. (2022). A study on the psychology of social engineering-based cyberattacks and existing countermeasures. *Applied Sciences, 12*(12), 6042. https://doi.org/10.3390/app12126042

SOCRadar. (2020, June 2). An overview of Verizon's 2020 data breach investigation report: A deep look into attacks, attackers and victims. *SOCRadar*. Retrieved from https://socradar.io/an-overview-of-verizons-2020-data-breach-investigation-report-a-deep-look-into-attacks-attackers-and-victims/

Steve, M. (2019). *Cybersecurity Ventures' 2019 Cybersecurity Market Report* . Sausalito, Calif: Morgan Steve Editor-in-Chief.

Syafitri, W., Shukur, Z., Mokhtar, U. A., Sulaiman, R., & Ibrahim, M. A. (2022). Social engineering attacks prevention: A systematic literature review. *IEEE Access, 10*, 39325–39343. https://doi.org/10.1109/ACCESS.2022.3162594

Tahir, Mirza, Amaad & Asghar, Sohail & Zafar, Ayesha & Gilani, Saira. (2016). A hybrid model to detect phishing-sites using supervised learning algorithms. *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 1126-11331126-1133. doi 10.1109/CSCI.2016.0214.

Tang, L., & Mahmoud, Q. H. (2021). A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction, 3*(3), 672-694.

Tushkanov, V. (2023, May 1). Investigating ChatGPT phishing detection capabilities. *Securelist*. https://securelist.com/chatgpt-anti-phishing/109590/

Uebelacker, S., & Quiel, S. (2014). The social engineering personality framework. *Proceedings of 2014 workshop on socio-technical aspects in security and trust*, 24–30. https://doi.org/10.1109/stast.2014.12

Venkatesha, S., Reddy, K. R., & Chandavarkar, B. R. (2021). Social engineering attacks during the COVID-19 pandemic. *SN computer science, 2*(2), 78. https://doi.org/10.1007/s42979-020-00443-1

Verizon. (2022). 2022 Data Breach Investigations Report. *Verizon Business*. Retrieved from https://www.verizon.com/business/en-gb/resources/2022-data-breach-investigations-report-dbir.pdf

Violino, B. (2023, January 7). Phishing attacks are increasing and getting more sophisticated. Here's how to avoid them. *CNBC*. Retrieved from: https://www.cnbc.com/2023/01/07/phishing-attacks-are-increasing-and-getting-more-sophisticated.html

Winder, D. (2019, Aug 20) Data breaches expose 4.1 billion records in first six months of 2019. *Forbes*. Retrieved from: https://www.forbes.com/sites/daveywinder/2019/08/20/data-breaches-expose-41-billion-records-in-first-six-months-of-2019/?sh=441acf5ebd54

Yeboah-Boateng, E. O. & Amanor, P. M. (2014). Phishing, smishing & vishing: An assessment of threats against mobile devices. *Journal of Emerging Trends in Computing and Information Sciences*, *5*(4), 297–307.

Zacharakos, A. (2023, February 2023). How hackers can abuse ChatGPT to create malware. *TechTarget.* Retrieved from: https://www.techtarget.com/searchsecurity/news/365531559/How-hackers-can-abuse-ChatGPT-to-create-malware

Zhang, X., Yadollahi, M. M., Dadkhah, S., Isah, H., Le, D. P., & Ghorbani, A. A. (2022). Data breach: analysis, countermeasures and challenges. *International Journal of Information and Computer Security, 1*9(3-4), 402-442.

Appendices

Appendix A

- Models are saved in: https://github.com/l-hawamdah/PhishNet_GWUPraxis

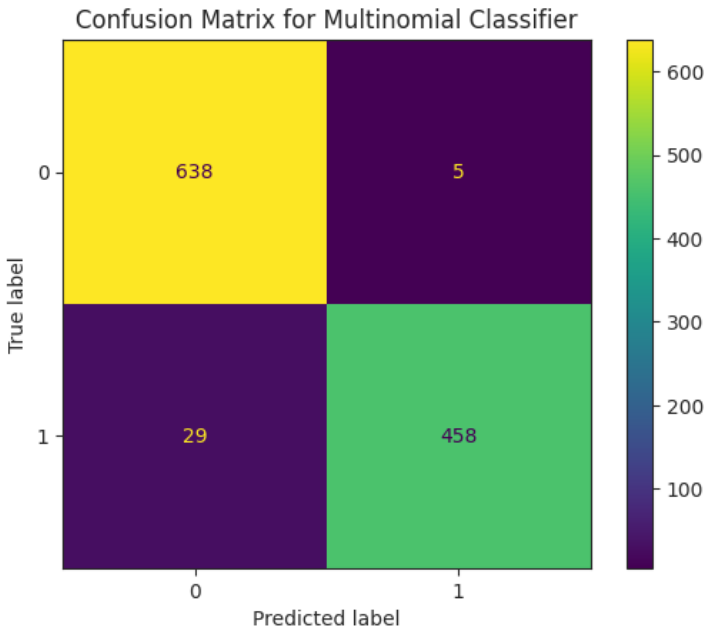


Figure A-1: Phishing Email Dataset Analysis on Multinominal

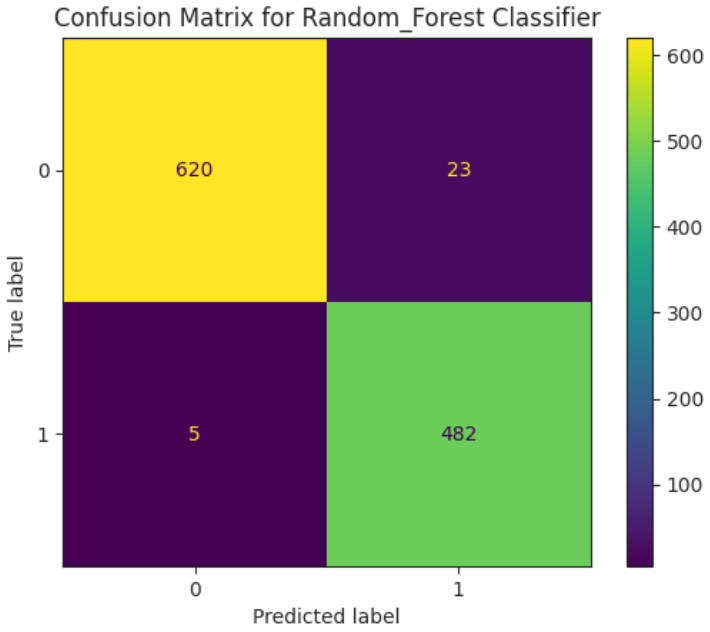


Figure A-2: Phishing Email Dataset Analysis on Random Forest

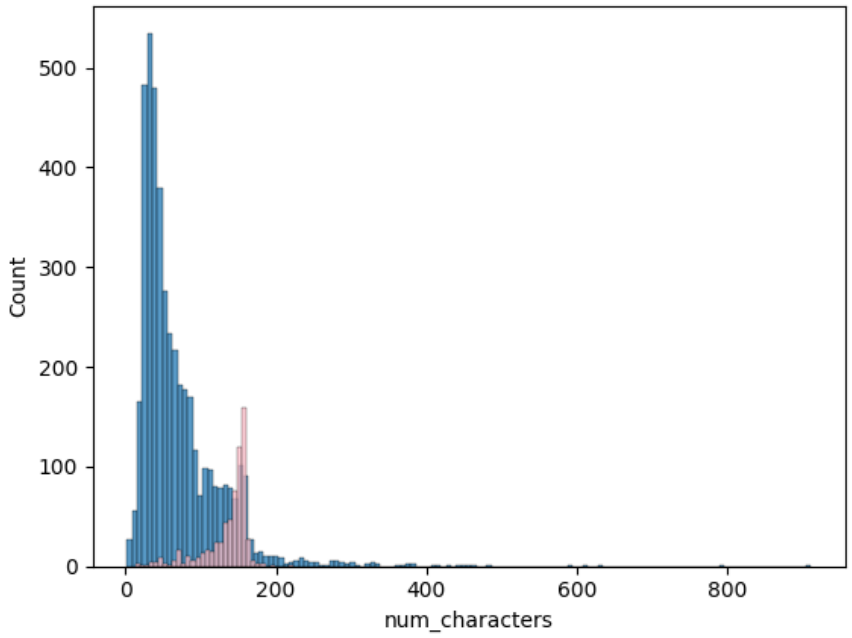


Figure A-3: Phishing Text Message Number of Character Distribution

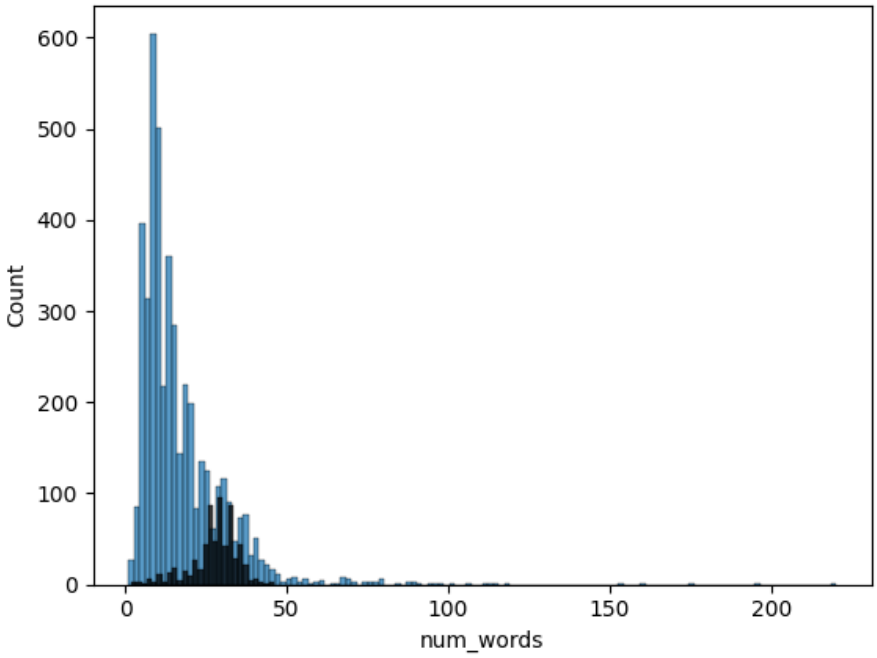


Figure A-4: Phishing Text Message Number of Words Distribution

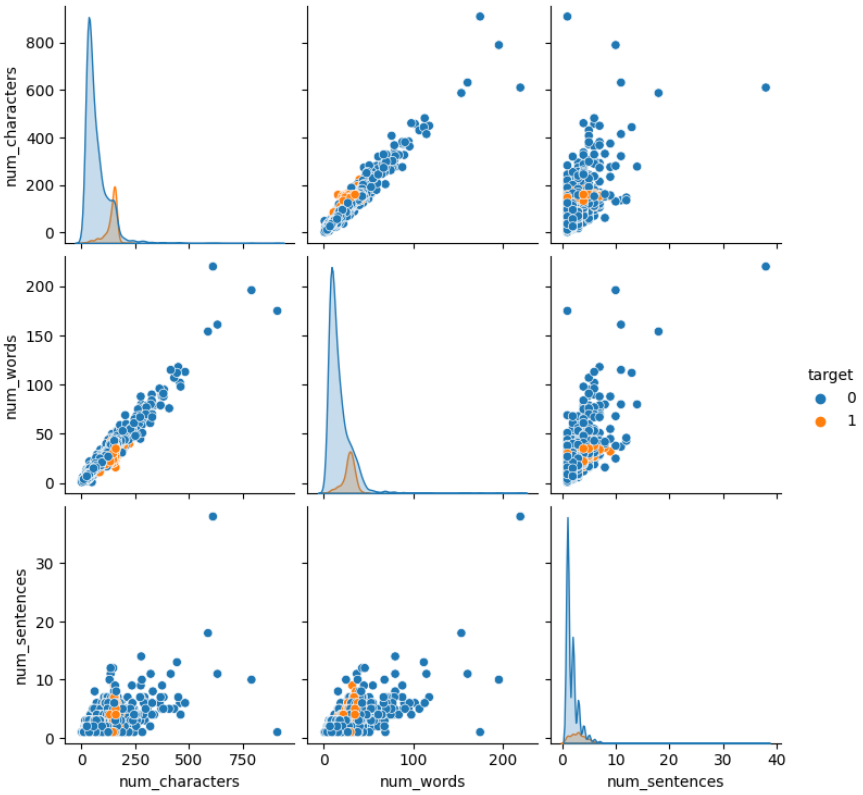


Figure A-5: Phishing Text Message Feature Relationship

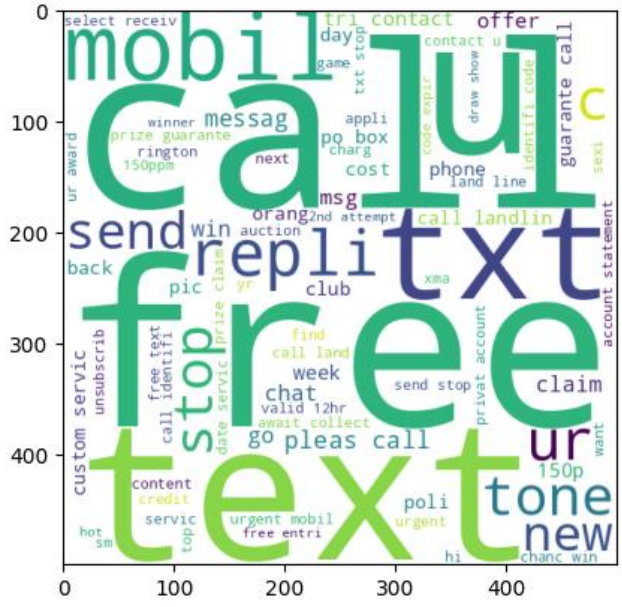


Figure A-6: Common Phishing Words in Phishing Text Dataset

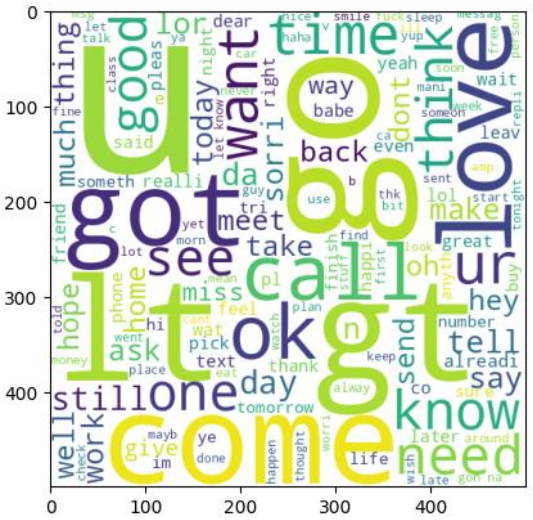


Figure A-7: Common Legit Words in Phishing Text Dataset

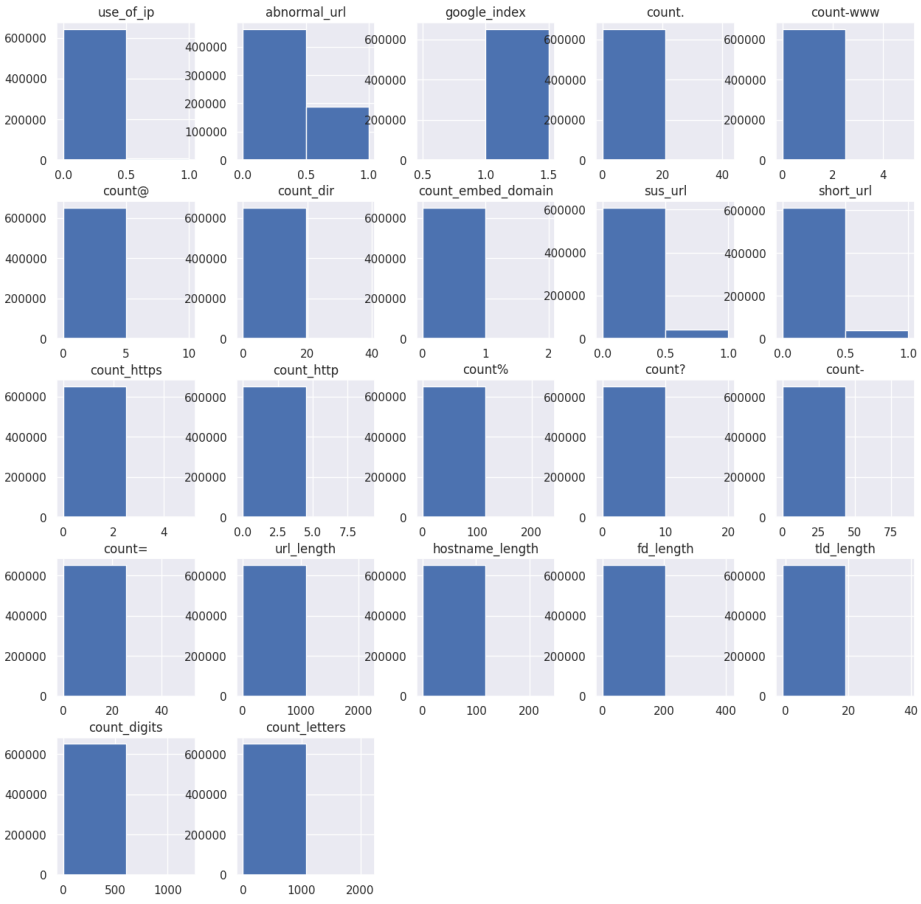


Figure A-8: URL Feature Engineering Distribution

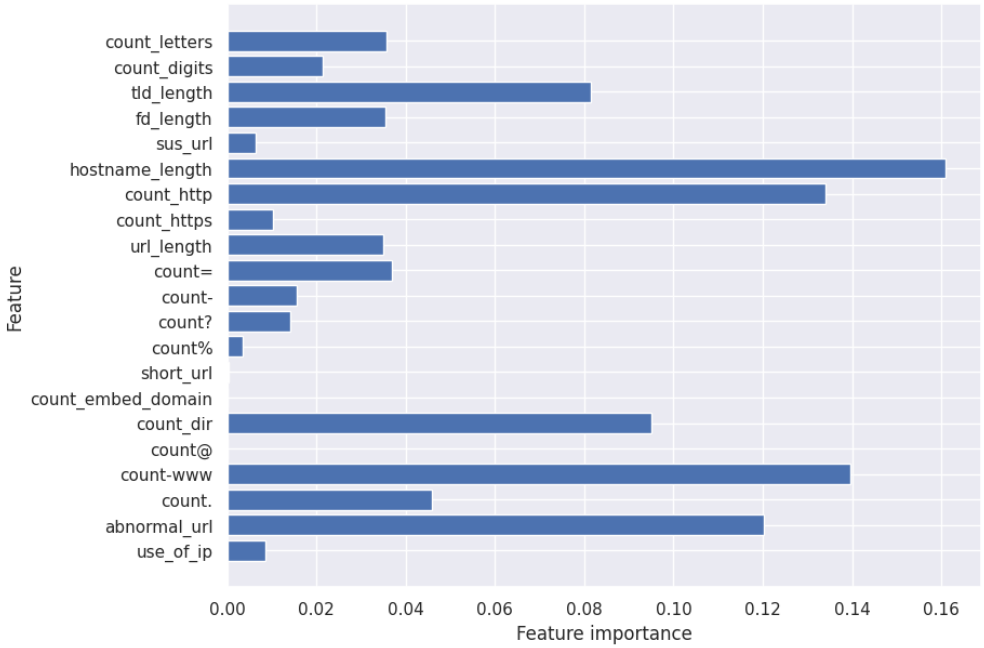


Figure A-9: URL Feature Engineering Feature Importance Distribution