Kehoe, Joseph

kINGSTON UNIVERSITY

**Machine Learning Approaches for Phishing Detection with Natural Language Processing**

Table of Contents

Introduction

1.1 The Evolution of Cyber Threats

1.2 The Role of AI and NLP in Phishing Detection

1.3 Structure of the Dissertation

The Dual Use of AI in Phishing Detection

2.1 Benefits of AI and NLP in Security

2.2 Risks and Challenges

2.3 Ethical Considerations

Survey of Existing Technology

3.1 Overview of Current Technologies

3.2 Case Studies

3.2.1 Microsoft's Approach

3.2.2 Google's Techniques

3.2.3 Behavioral Interventions at EE and BT

3.3 Comparative Analysis

User Education and Awareness

4.1 Importance of User Education

4.2 Identifying Phishing Attacks

4.3 Prevention and Best Practices

4.4 Reporting Phishing Attempts

Developing a Phishing Detection Tool

5.1 Project Overview

5.2 Model Configuration and Implementation

5.3 Dataset Challenges and Solutions

5.4 Performance Evaluation and Results

5.5 Comparison of Tools: Python Script vs. Email Guardian

Conclusion

6.1 Summary of Findings

6.2 Future Research Directions

6.3 Ethical and Dual Use Consideration

References

Appendix

Introduction:

In the digital age of today, the proliferation of cyber threats continues to pose significant challenges for organizations and individuals worldwide. Among these threats, phishing attacks are one of the most common and damaging forms of cybersecurity threats.

**Phishing** is the practice of sending fraudulent communications, typically via email, that appear to come from a reputable source with the intent to steal sensitive information, such as credit card information. As these attacks have grown more sophisticated, leveraging on human error and the subtle phycology that makes someone trust a link, the need for advanced deference measures has become paramount.

This paper will explore the critical role of Artificial Intelligence and Natural Language processing in phishing detection. The integration of AI and NLP technologies has revolutionized the way we detect and prevent phishing attacks, providing tools that can analyze huge amounts of data at speeds and accuracy far beyond human capacity. These technologies are essential for identifying phishing hallmarks but also for adapting to the continuously evolving tactics employed by cybercriminals.

**Natural Language Processing** (NLP) is a branch of AI that focuses on allowing computers to understand, interpret and produce human language in a way that is valuable and meaningful. NLP combines computational linguistics with (rule based modern language) with machine learning and deep learning models. These technologies allow systems to process human language in the form of text to understand its full meaning, including the intent and sentiment.

Natural Language processing also plays a pivotal role in deciphering the textual nuances and intent within emails. By employing techniques such as sentiment analysis, contextual relevance analysis and semantic text evaluation NLP helps distinguish malicious emails from legitimate ones. This is crucial in identifying more sophisticated phishing attacks that may not contain suspicious elements like links but rely on manipulating the recipient through persuasive language.

This trend of relying on AI and NLP in phishing detection reflects a trend towards automated, intelligent cybersecurity measures. As techniques for phishing become more refined, the intersection of AI and NLP represents a promising future in the ongoing battle against cyber threats.

In addition to exploring established AI and NLP techniques used by leading firms, this dissertation also showcases a personal endeavor in the field: the development of an AI tool designed to combat phishing threats. This project represents an integration of theoretical knowledge and practical application.

The Dual Use of AI in phishing

While AI and NLP offer significant advantages to the detection and prevention of phishing attacks it is crucial to acknowledge the dual edged nature of these technologies. AI capabilities can also be leveraged by malicious cybercriminals to craft and distribute more sophisticated phishing emails. The paradoxical use of AI highlights a critical challenge in cybersecurity; as defensive technology evolves, so do the methods of attackers.

AI tools can automate the generation of phishing content by creating personalized and contextually relevant messages that would mimic legitimate communication. For example, machine learning models can analyze vast amounts of data from social media and other public sources to tailor deceptive emails that resonate more personally with potential victims. This customization increases the chance of users engaging with these malicious emails as they would appear more credible and relevant.

NLP technology enhances the believability of phishing emails by improving the linguistic quality of the phishing content. It can be used to generate grammatically correct, stylistically appropriate messages that can accurately imitate the tone and vocabulary of a legitimate communication from a trusted entity. This capability makes it increasingly difficult for users and phishing detection systems to distinguish between legitimate and fake messages.

The use of AI for creating phishing emails shows a significant shift in the landscape of cyber threats, requiring more advanced AI enhanced defensive strategies. It highlights the importance of developing AI systems that not only detect phishing based on static indicators, but also the ability to learn and respond to new and emerging tactics employed by cybercriminals. This “Arms Race” in AI capabilities between cyber security professionals and attackers drives the need for continued research, development and ethical considerations in the use of AI technologies.

When addressing the role of AI and NLP in phishing detection, it is essential that both positive impacts and the potential misuse of these technologies. By understanding how AI can be employed by attackers, cyber security professionals can better foresee future threats and increase their defensive measures against an ever-adaptive enemy.

Survey of existing technology

This section will present a comprehensive survey of the technologies employed by leading technology companies to detect and mitigate phishing emails. Given the importance of phishing threats, which exploit human factors as much as technical vulnerabilities, companies invest heavily into sophisticated detection systems that blend traditional security methods with robust artificial intelligence and machine learning techniques. This survey encompasses a selection of notable companies, recognized for their contributions to email security. The focus lies on the diverse array of approaches and technologies such as Natural Language processing, machine learning classifiers, contextual analysis and more. These technologies are never isolated but part of an integrated approach to predict and prevent phishing attempts.

It is important to acknowledge that there may be redundancies in the information presented, as these companies often employ overlapping technologies and methods. For example, the use of machine learning is a common theme across most phishing detection systems, as well as NLP and feature selection.

Microsoft:

Microsoft employs a multi layered approach to email phishing detection, using both proprietary technology and partnerships with cybersecurity firms. This comprehensive strategy is in place to protect users from a wide range of email-based threats, including phishing, spear phishing and malware attacks.

1. Advanced Machine Learning Models

Microsoft uses cutting-edge machine learning models that analyze email content to detect phishing and spam attempts. These models are trained on vast datasets, allowing them to identify patterns and characteristics in common phishing emails, even as attack strategy evolves. Microsoft also collaborates with firms like SlashNet to enhance its detection capabilities. SlashNet’s AI based detection service is particularly effective against spear phishing and more advanced threats with a high degree of accuracy and minimal false positives.

1. Anti Malware Protection

Emails are scanned for malicious attachments and URLS. Signature-based detection engines are used to compare files and links against a database of known malware signatures, allowing for the prompt identification and neutralization of threats.

1. Domain and IP Reputation Systems

BY analyzing the reputation of IP address domains from which emails originate, Microsoft can identify and block emails that come from known malicious sources. This includes analyzing the history of a domain’s behavior and involvement in sending spam or malicious content.

1. Sender and Domain Impersonation Detection

Microsoft's technology can identify attempts to impersonate legitimate brands domains and users. This provides protection against domain spoofing where attackers mimic a legitimate domain to trick users into believing the origin is a trusted source.

1. URL Analysis and Detonation

URLs within emails are also checked against databases of known links. Microsoft also uses URL detonation techniques where links are opened in a controlled virtual environment to observe their behavior, identifying and blocking harmful URLS before they can reach the user.

1. Mailbox Intelligence

There is a personalized protection mechanism in place for each user that builds a map of trusted senders based on habits. This helps in detecting impersonation attempts by comparing incoming emails against this map, largely reducing the likelihood of successful phishing attempts.

1. Bulk Complaint Level (BCL)

This metric evaluates how likely an email is to be spam based on how often emails from the same source are marked as spam by users, helping to filter out mass spam and phishing campaigns.

Google

1. Natural Language processing
2. Feature Selection and Classifier Creation

Google applies statistical tests to analyze email content, it identifies distincitive features that are more prevalent in phishing emails than legitimate ones. These features include certain phrases, the presence of URLS, and stylistic elements. Once Identified these features are used to train classifiers; which are algorithms that can sort incoming messages into “phishing” or legitimate based on their content.

1. Contextual information

Google also checks the context of the users email history and interactions. This analysis helps in identifying discrepancies or anomalies in incoming emails, such as from senders that the user has not interacted with before or emails that deviate from the typical communication patterns observed in the users account. This method has been shown to be effective against more sophisticated phishing attempts that mimic legitimate users.

1. Visual-AI (Computer Vision)

Google uses computer vision technology to inspect the visual components of an email. This includes analyzing embedded images and stylized text that may be used to disguise malicious links or deceptive calls to action. By rendering the content as it would appear in a web browser, Google’s system can adequately detect attempts to visually clicking a phishing link.

1. Deep Learning Algorithms

Deep learning forms a backbone of Googles phishing detection system, with models trained on huge datasets of both phi shing and legitimate emails. These models learn to identify subtle patterns and indicators that may not be visible to human reviewers or traditional algorithms, ensuring the detection of even the most sophisticated phishing schemes.

1. Machine Learning Model Deployment at a scale

To handle the vast volume of emails processed every day, Google has developed a scalable system for deploying machine learning models. This includes the continous monitoring for model performance, regular updates to the training datasets to include new threats and the use of systmes like “Datadog” to moniro model performance in real time. This ensures models stay effective over time and are able to adapt to new phishing strategies as they emerge.

1. Comprehensive Threat Intelligence

Googles anti phishing techniques are integrated with broader threat intelligence systems that collect data from millions of user interactions daily. This global perspective allows google to quickly identify and respond to emerging phishing trends, provind a level of protection that is both deep and wide ranging.

Behavioral Interventions at EE and BT)

As part of this survey on phishing detection, it is pertinent to include my firsthand insights obtained from my professional experience and EE and BT. My position within these companies has given me the unique opportunity to access internal strategies and firsthand data regarding cybersecurity practices. Which includes a proactive approach to educating employees.

BT uses external partners like Microsoft for email protection but also employs complementary strategies focusing on cybersecurity. One notable tactic includes the intentional sending of simulated phishing emails to employees. This proactive educational approach has a critical function in enhancing organization security.

This strategy's main goal is to educate and train employees on identifying and handling phishing attempts. By exposing employees to controlled examples, BT and EE aim to create a heightened awareness of phishing attacks. Employees learn to recognize the hallmarks of phishing emails which increases their ability to discern legitimate emails from those with malicious intent.

There are also multiple mandatory courses on phishing emails and how to detect them in which employees interact with emails to differentiate them from phishing and legitimate. This allows them to recognize similar threats in the future and understand the correct response to these emails, such as the reporting and blocking of suspicious content.

Feedback from these exercises is used to continuously improve the training process, and from this the simulations can be tailored to reflect the latest phishing techniques and trends, ensuring that the training remains relevant and effective.

This strategy contributes to cultivating a strong cybersecurity culture within the business. By regularly engaging employees in these exercises, EE and BT reinforce the critical role of everyone in cyber security defenses. This not only enhances the ability of employees to deal with phishing attacks but also promotes a security conscious organization environment.

User Education:

User education is pivotal in combatting phishing attacks. Simply put, the more a user understands what a phishing email is and how to identify it, the more likely they are to safeguard their data. The following is a standalone user guide, it may repeat some information from earlier in the paper.

**Phishing Education Guide**

*Today is a digital age, with our lives intertwined with the internet. Nine in ten Americans are using some form of digital payment(*[*Lindsay Anan*](https://www.mckinsey.com/industries/financial-services/our-insights/banking-matters#/author/lindsay-anan)*), and phone based wallets such as google and apple pay are rapidly rising in popularity. With these facts in mind, it is obvious that the security of our digital information could never have been more paramount. Amongst this overwhelming surge of online data lies a ever growing threat; Phishing.*

*So what is Phishing?*

*Phishing is a genre of cybercrime where attackers will use social engineering to deceive individuals into revealing sensitive information that will grant them access to their data. Social engineering simply refers to techniques that are used for manipulating individuals, and it can take many forms; Phishing being the most common and dangerous.*

*Phishing attacks can come in many forms, such as deceptive emails, fraudulent websites and misleading texts. They are all designed to appear as genuine communication from trusted entities. These attacks use trust and the naivety of individuals to achieve a generally malicious objective, ranging from identity theft to the much more common financial fraud, and in some cases gain access to a large secure network.*

*Recognizing Phishing Attacks:*

*Phishing messages often contain telltale signs that can cause alarm to a cautious user. Here are some key indicators.*

*Suspicious Sender Addresses: Always check the senders email address carefully. Phishing attempts often come from addresses that will attempt to mimic a legitimate one, with small changes such as letters or numbers.*

*Urgency in the Message: Attackers often attempt to create a sense of urgency to provoke a quick response. Be wary of emails or messages that pressure immediate action, especially when concerning personal details or payment information.*

*Grammar and Spelling Errors: Legitimate messages from actual organizations are usually well written. Poor grammar and spelling mistakes are a key indicator of a phishing attempt.*

*Unusual Requests: Be suspicious of requests for your information, such as passwords, credit card numbers and more so when these requests come out of the blue. Companies will not usually request this sort of information, especially over email correspondence.*

*Mismatched URL: Hover over any links in the email before clicking them to check fi the actual URL looks different from the hyperlink text; it could indicate a phishing attempt.*

*Attachments: Be extremely cautious with email attachments, as these can contain malware. Only open attachments you are expecting and from senders you trust.*

*Types of Phishing Attack:*

*Phishing attacks have been designed to extract specific types of information or deceive groups of people. Understanding these will make you better equipped to recognize attacks.*

*Spear Phishing: Spear Phishing targets specific individuals or organizations. Attackers gather personal information about their targets to create highly customized messages. You should always verify the information in unexpected requests, even when they appear to be from within your own organization or from a known contact.*

*Whaling: Whaling attacks target seniors and other high-profile targets within businesses. The content of these emails can be pressing issues such as legal subpoenas or customer complaints.*

*Vishing (Voice Phishing): In these attacks' phishers use phone calls to extract personal information. They may impersonate banks, government officials or other authorities to seem legitimate.*

*By keeping an eye out for these indicators and understanding the types of phishing attacks, you can enhance your own defense against this common cybercrime.*

*Prevention and Best Practice:*

*The prevention of phishing attacks is a combination of technical safeguards and cautious behavior. Educating yourself and others on these best practices is crucial for protecting your sensitive information and maintaining security.*

*Security Measures:*

*Spam filters: Use spam filters to block suspected phishing emails from reaching your inbox. Most modern email systems have filters pre-installed.*

*Secure connections: Ensure the websites you visit use HTTPS, especially if you are entering private information. This is a protocol that secures the data transfer between the browser and the server.*

*Software updates: Keep all your software up to date, including your operating system, browser and any other apps you use. Updates often contain security updates to protect against new threats.*

*Two-Factor Authentication: Enable 2FA on all accounts that offer it. This adds an extra layer of security by requiring a second form of verification in addition to a password.*

*Behavioral Changes:*

*Verify Unexpected requests: if you receive an email that requests sensitive information, you can verify it by contacting the company directly using a known email or phone number, not the contact provided within the message.*

*Think Before You Click: avoid clicking links or downloading attachments from unknown sources. If you must, always verify the link by hovering over it first.*

*Regular Monitoring: Regularly checking bank statements, credit reports and account summaries for unusual activity can help early detection of successful phishing attempts, mitigating damage caused.*

*Reporting Phishing Attempts:*

*Knowing how to report phishing attacks can help reduce them.*

*Internal Reporting: If you’re part of an organization, report suspected emails to your IT or security team.*

*External Reporting: Most email services will allow you to flag and report and email as malicious, you can also report more sophisticated phishing attempts to relevant authorities. You can also forward fraudulent emails to the Anti-Phishing Working group at reportphishing@apwg.org.*

*Example Analysis: Spotting an Attack*

*A screenshot of a computer

Description automatically generated*

*Senders Address: check the senders address by looking at the domain after the @. it may resemble a real companies address but with subtle changes, in this case paypal-accounts.com has been used to pass as an official PayPal email, which would generally come from paypal.com.*

*Urgency and Threats: This email creates a state of emergency stating the account will be “permanently disabled” within 24 hours. This is a common phishing technique used to hurry the user into taking action before they have time to critically analyze the email.*

*Grammar and Spelling: Look for unusual grammar or spelling, mistakes a professional company usually would not make. There are a few grammatical errors and capitalization mistakes in the above email.*

*Link to a website: Typically, there is a call to action, such as the button presented here. Before clicking hover over the link to see if the URL looks suspicious.*

*Branding: Often phishing emails will use logos and brandings to appear legitimate. However closer inspection may show discrepancies.*

Developing a Phishing Detection Tool

Python Script

The project is centered on the integration of GPT 3.5 Turbo, a cutting-edge language model developed by OPENAI. The choice of GPT-3.5 turbo leverages its powerful natural language processing capabilities, which takes away the need for a huge amount of training of the model. This can analyze and interpret contents of emails more efficiently than conventional methods. The original project plan; which was to train an AI on an email dataset from real people, was more ambitious than I expected, and the reasons and drawbacks for this will be explained, as well as ways this could be achieved in future with the correct budget.

By utilizing this AI approach this project aims to:

Detect subtle and sophisticated phishing cues that are often overlooked by standard phishing filters that focus on blatant indicators.

Automate the detection process with greater accuracy and efficiency, reducing the associated human error.

Adapt and respond to new phishing strategies as they emerge, using the models' learning capabilities to update its detection algorithms based on new data.

The importance of using **GPT-3.5 turbo** within the program lies on its ability to understand and process natural language at a level that mimics human comprehension, but at a scale and speed that could not otherwise be achieved. This project will explore how advanced NLP can be a huge step in phishing detection, providing a robust tool against a prevalent cyber threat. GPT 3.5 was selected primarily on its natural language processing capabilities, it’s ability to interpret nuanced language in text hugely important in phishing detection. GPT-3.5 turbo is a very powerful engine, with a learning algorithm that is constantly learning from new data making it an adaptable tool. It is also incredibly efficient at processing data which makes it highly suitable for phishing detection. Semantic analysis allows detailed understanding of words in various contexts, a vital point for distinguishing between legitimate and fraudulent messages.

1. Model Configuration:

The project utilizes the OpenAI API in order to access 3.5 turbo. The API calls are configured to pass the email content to the model and then receive an assessment of potential phishing activity based on the textual analysis.

Custom Prompts to tailor the model’s responses for phishing detection, custom prompts were used to instruct GPT 3.5 turbo on the specific task. These prompts guide the model to focus on identifying key phishing indicators within the emails, such as deceptive language, misleading links, or requests for sensitive information.

Data Preprocessing:

Text Extraction: Emails are often composed of complex structures including headers and footers. A preprocessing application was used to extract only relevant text from each email.

Normalization: The text data was then manually normalized to reduce variability, to make the model's job of understanding and deciphering the text easier.

1. Dataset Drawbacks:

Creating a robust database for testing and training a machine learning model, involves overcoming significant challenges. In this project, several obstacles impacted the ability to compile a comprehensive dataset of real-world emails.

Low Response Rate: The initial aim to gather a wide range of emails from real world individuals to create a dataset that accurately represents the diversity of email communication. However, the response rate was much lower than originally anticipated. Many individuals were hesitant to share their emails due to privacy concerns or disinterest, resulting in a dataset that was not large enough to comprehensively test or train a model.

Even among the dataset collected, mainly made up of emails from online sources, the diversity of email types was limited, and this lack of variety could potentially bias the model towards specific types of emails or phishing attempts, which could reduce the overall effectiveness of the detection system.

High Cost of training: while GPT 3.5 turbo is a powerful tool, it comes with a significant cost especially when processing large datasets. OpenAI charges based on the tokens processed which can quicky escalate the costs when training on extensive collections of emails. Budget constraints therefore severely limited me in terms of training. Without adequate funding it was not feasible to conduct extensive model training as intended, which may have yielded better results.

1. Potential Solutions:

Improving the response rate for email collection through better outreach strategies and incentives for users could help in gathering a more extensive and diverse dataset.

Collaborating with organizations that already have access to large and diverse datasets could alleviate both the data and cost issues. Sponsorship from cybersecurity companies interested in advancing phishing detection technology would provide the necessary funding to cover the high costs of training GPT 3.5 turbo.

1. Choice of Python:

Python was chosen for the coding of the program for several compelling reasons.

Run Library Ecosystem: Python supports many libraries that facilitate machine learning and data processing tasks. Libraries like ‘openai’ for accessing OpenAI’s API, along with other for data manipulation such as “pandas”.

Simplicity and Readability: Python has a clear and intuitive syntax, making it accessible for both seasoned programmers and novice programmers. Readability is important for projects including machine learning as ease of learning for maintenance is important.

Community Support: Python has a large and active community, which brings a wealth of tutorials and support. The basics of programming with GPT 3.5 were readily available.

These reasons and the easy integration of other services into python made it the obvious choice.

1. Code explanation:

The code for the project revolves around using the OpenAI API to assess emails for phishing risks. Here is a small breakdown of the code created.

A screenshot of a computer program

Description automatically generated“import open ai”

This line imports the openai library which is necessary to interact with GPT. The library provides functions to make requests to the gpt-3.5 turbo model and handle responses.

The next step was to import my GPT API key. To configure this in python we simply use.

API\_KEY = "KEY GOES HERE"

The function “read\_mail” reads the content of an email from a specific file path.

It also includes error handling using a try-except block for cases where the specified file does not exist, which prevents the script from crashing offering a more user-friendly experience.

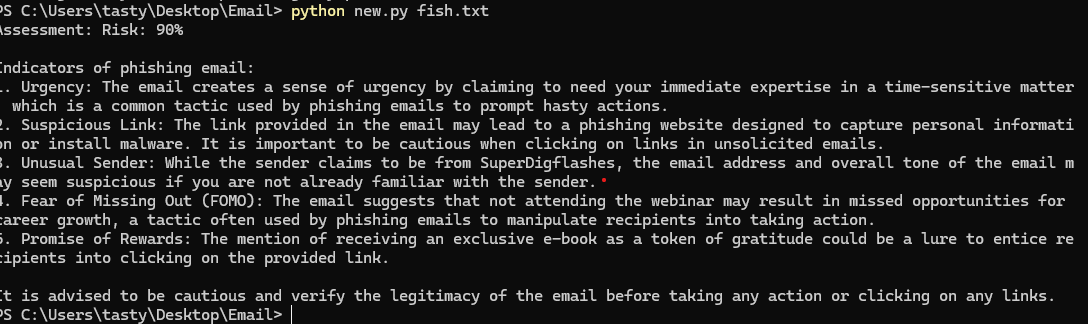
The function ‘create\_prompt’ formats the email context into a prompt, that is then sent to the OpenAI API, instructing it to analyze the email for phishing risks, and quantifying the risk into a percentage.

The function ‘query\_openai’ is used to send the newly crafted prompt to OpenAi using the GPT-3.5 Turbo model and handles the response. It also has error handling that captures issues during the API request, such as connectivity issues or API limtis, that were experienced many times during the project.

Main Code Block

The script will accept the file path of the email as a command line argument, which enhances the usability and flexibility for batch processing or the integration with larger systems. The process flow runs the previous functions, reading an email, creating a prompt, sends to the GPT API and model then prints the phishing risk assessment.

The result will output like this.



As demonstrated, the script effectively uses the GPT 3.5 turbo model and its natural language processing powers to analyze potential phishing emails and explain its reasoning behind the risk assessment. This shows the practical application of AI and natural language processing in cybersecurity through python programming.

Email Guardian:

The GPT creator tool offered by OpenAI also presents a significant leap forward in phishing detection capabilities in compared to the python script solution. GPT allows you to configure and build a model of 3.5 within its own ecosystem. A version was configured to use NLP and common phishing detection methods to successfully detect phishing attempts and explain them and named “Email Guardian”. This tool enables the analysis of email content in various formats, including the ability to process screenshots which adds a crucial layer that textual analysis alone may miss. Some advantages of this tool are:

Multimodal Analysis: The most crucial feature of this tool, named “Email Guardian” is its ability to not only analyze plain text but visual information present in screenshots. This is especially important considering that many phishing attempts using images to convey information would avoid text-based detection systems.

Mobile usability: With the use of mobile phones at an all-time high, phishing attempts will often target mobile users. “Email Guardian” enhances usability by allowing users to simply take a screenshot, then upload it within the GPT app, which is a significantly easier method when compared with more traditional techniques.

Real Time analysis: Users can receive real time feedback on potential risks directly within their mobile environment, making it a highly accessible and immediate solution for email security.

Comparison with Python Script Program:

While the python scripted program is a powerful tool that is capable of scaling to analyze large volumes of textual data effectively, it does have limitations that “Email Guardian” addresses.

Text only analysis: The python script is limited to text, meaning it cannot detect phishing attempts embedded in images or elaborate html designs that are becoming more common in sophisticated phishing emails.

Cost considerations: While both “Email Guardian” brings enhanced capabilities to phishing detection, particularly its ability to analyze visual content and its suitability for mobile use, making it an excellent tool for individual users and on the go analysis. However, the python scripted program with its focus on text analysis and scalability, is more suited for enterprise-level solutions where large volumes of emails need to be checked and processed.

Both tools provide contextual information about the emails, and indicators of phishing explained in a user-friendly way. They each have unique strengths and choosing between them should depend on the specific needs of the use case.

Performance Evaluation:

Test setup: For the performance evaluation 4 phishing emails and 4 legitimate emails were tested. The emails were sorted into readable text for the python program to understand. Those same emails will be used in the testing “Email Guardian” however they will be presented as screenshots where possible. The dataset for evaluating the performance of the tool may seem small, again this is due to budgetary constraints. The results of both tests will be available for viewing, the python program as a video and “Email Guardian” as a text document.

Python results:

Out of the emails tested, the python program detected all the phishing emails. Giving a risk assessment of 90%, 100%, 100% and 100%. It also provides reasoning for each decision. However, out of the real emails tested there were some false positives. Giving a risk assessment of 50%, 5%, 40% and 70% for legitimate emails. It does, however, give reasoning for why users may proceed with caution. The reasoning for this result is likely due to the text-based nature of the detection system, the lack of ability to read html style and embedded images leaves a large loophole for more sophisticated phishing attempts.

Email Guardian results:

Email Guardian can process screenshots, so where possible the emails were provided as images so additional elements can be detected. While explaining how and why it came to these conclusions it ranked phishing emails, 90%, 90%, 50-70% and 70-80%, proving itself as a capable tool that can check for embeds. There was a significant decrease in false positives within this program, ranking the genuine emails: 10%, 30%, 10-20% and 10%. This is likely due to the visual context available within screenshots that the program can analyze. Context and reasoning are also provided from the program, which adds to its usability. Email Guardian was also able to process all these emails in one query. While this may still be possible in the python script, more work is required when compared with this tool.

Conclusion

This project has explored the critical role of AI and NLP in forwarding phishing detection techniques. By integrating advanced techniques, this work has demonstrated potential improvements in the identification of sophisticated phishing attempts that traditional techniques might lack. The development of both Email Guardian and the use of a Python-based Ai model highlight the potential of using AI to enhance cybersecurity practices by offering:

Enhanced Detection Accuracy: Both programs have shown high effectiveness of detecting phishing attempts using NLP, with “Email Guardian” providing additional capabilities through visual analysis.

Reduction of False positives: Particularly with the development of “Email Guardian” the number of false positives has been reduced, thanks to multimodal analysis, which is a common problem in phishing detection.

The implementation of AI and NLP in phishing detection addresses current threats and adapts to evolving tactics used by cybercriminals. This adaptability is crucial as more sophisticated and well-crafted phishing attacks emerge requiring more proactive defenses. The research provided contributes to a better understanding of how AI tools can be developed to secure digital communications against emerging cyber threats.

Ethical and Dual Use Consideration:

This work has also highlighted the dual use nature of AI technology in cybersecurity. While AI offers a powerful set of tools for defense, its capabilities can be equally misused by malicious individuals or groups. It is important that ongoing research into AI is guided by ethical standards, the development and deployment of these technologies should be monitored and analyzed to ensure the overall impact of AI and NLP is positive.

Further Research Directions:

Data Diversity and Model Training: To improve the python AI model, future work should focus on a diverse dataset for training as well as funding for implementation into gpt3.5-turbo to cover a broader spectrum of phishing techniques.

Real-Time Phishing Detection systems: Integrating models such as the python script and Email Guardian into real-time systems could provide instant protection, adapting to threats as they grow and occur.

Cross Platform Solutions: As phishing attempts have grown to attack across various digital platforms, developing a solution that works with AI cross-platform could provide a more comprehensive security solution.

In conclusion, the integration of AI and NLP into phishing detection represents a promising future in the ongoing battle against cyber threats. As the field grows, it will require continuous research and collaborations between programmers and policy makers to effectively use AI to its full potential, while maintaining its responsibility. The trajectory of AI development shows a promising future for digital security, where users can be safeguarded from threats before they can manifest.

References)

Phishing.org, n.d. Phishing Techniques. Available at: https://www.phishing.org/phishing-techniques [Accessed 12 Jan 2024].

PhishProtection, 2019. Phishing Attack Prevention. Available at: https://www.phishprotection.com/content/phishing-prevention/phishing-attack-prevention [Accessed 12 Jan 2024].

Proofpoint, n.d. Phishing. Available at: https://www.proofpoint.com/uk/threat-reference/phishing [Accessed 13 Jan 2024].

Perception Point, n.d. How to Prevent Phishing Attacks. Available at: https://perception-point.io/guides/phishing/how-to-prevent-phishing-attacks/ [Accessed 13 Jan 2024].

Wikipedia, 2024. Phishing. Available at: https://en.wikipedia.org/wiki/Phishing [Accessed 13 Jan 2024].

BlueVoyant, n.d. 8 Phishing Types and How to Prevent Them. Available at: https://www.bluevoyant.com/knowledge-center/8-phishing-types-and-how-to-prevent-them [Accessed 16 Jan 2024].

Cofense, n.d. Anti-Phishing Best Practices. Available at: https://cofense.com/knowledge-center/anti-phishing-best-practices/ [Accessed 16 Jan 2024].

IT Governance, n.d. Phishing. Available at: https://www.itgovernance.co.uk/phishing [Accessed 18 Jan 2024].

Phishing.org, n.d. What is Phishing? Available at: https://www.phishing.org/what-is-phishing [Accessed 18 Jan 2024].

Kaspersky, n.d. Phishing Prevention Tips. Available at: https://www.kaspersky.com/resource-center/preemptive-safety/phishing-prevention-tips [Accessed 18 Jan 2024].

IT Governance EU, n.d. The 5 Most Common Types of Phishing Attack. Available at: https://www.itgovernance.eu/blog/en/the-5-most-common-types-of-phishing-attack [Accessed 18 Jan 2024].

CrowdStrike, n.d. Phishing. Available at: https://www.crowdstrike.com/cybersecurity-101/phishing/ [Accessed 18 Jan 2024].

TechTarget, n.d. Phishing. Available at: https://www.techtarget.com/searchsecurity/definition/phishing [Accessed 22 Jan 2024].

NIST, n.d. Phishing. Available at: https://csrc.nist.gov/glossary/term/phishing [Accessed 22 Jan 2024].

Phishing.org, n.d. 10 Ways to Avoid Phishing Scams. Available at: https://www.phishing.org/10-ways-to-avoid-phishing-scams [Accessed 22 Jan 2024].

Fortinet, n.d. Types of Phishing Attacks. Available at: https://www.fortinet.com/uk/resources/cyberglossary/types-of-phishing-attacks [Accessed 22 Jan 2024].

Trend Micro, n.d. Types of Phishing. Available at: https://www.trendmicro.com/en\_us/what-is/phishing/types-of-phishing.html [Accessed 26 Jan 2024].

Imperva, n.d. Phishing Attack Scam. Available at: https://www.imperva.com/learn/application-security/phishing-attack-scam/ [Accessed 26 Jan 2024].

NCSC, n.d. Phishing Scams. Available at: https://www.ncsc.gov.uk/collection/phishing-scams [Accessed 26 Jan 2024].

University of Salford, 2021. Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey Available at: https://salford-repository.worktribe.com/OutputFile/1563761 [Accessed 27 Jan 2024].

Springer, 2012. Detecting Phishing Emails the Natural Language Way. Available at: https://link.springer.com/chapter/10.1007/978-3-642-33167-1\_47 [Accessed 28 Jan 2024].

CyberSaint, n.d. AI in Cybersecurity. Available at: https://www.cybersaint.io/blog/ai-cybersecurity [Accessed 28 Jan 2024].

Aii.et, n.d. Why to Apply Natural Language Processing to Cybersecurity. Available at: https://www.aii.et/why-to-apply-natural-language-processing-to-cybersecurity/ [Accessed 28 Jan 2024].

VentureBeat, n.d. How to Apply Natural Language Processing to Cybersecurity. Available at: https://venturebeat.com/ai/how-to-apply-natural-language-processing-to-cybersecurity/ [Accessed 28 Jan 2024].

NVIDIA Developer Blog, n.d. Changing Cybersecurity with Natural Language Processing. Available at: https://developer.nvidia.com/blog/changing-cybersecurity-with-natural-language-processing/ [28 Jan].

Southern Methodist University, n.d. Phishing Detection Using Natural Language Processing and Phishing Detection Using Natural Language Processing and Machine Learning . Available at: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1215&context=datasciencereview [Accessed 28 Jan 2024].

Masernet, n.d. Role and Applications of NLP in Cybersecurity. Available at: https://www.masernet.com/project/role-and-applications-of-nlp-in-cybersecurity [Accessed 1 Feb 2024].

University of Salford, n.d. A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques. Available at: https://salford-repository.worktribe.com/OutputFile/1562817 [Accessed 16 April 2024].

Vertex Cybersecurity, n.d. AI Will Increase Phishing Cyber Attacks. Available at: https://www.vertexcybersecurity.com.au/ai-will-increase-phishing-cyber-attacks/ [Accessed 1 Feb 2024].

Spambrella, n.d. How is AI Enabling Phishing?. Available at: https://www.spambrella.com/how-is-ai-enabling-phishing/ [Accessed 11 Feb 2024].

Wired, 2021. AI is Making Phishing Emails Tougher to Spot. Available at: https://www.wired.com/story/ai-phishing-emails/ [Accessed 11 Feb 2024].

MetaCompliance, n.d. How AI Enables Sophisticated Phishing Attacks. Available at: https://www.metacompliance.com/blog/phishing-and-ransomware/how-ai-enables-sophisticated-phishing-attacks [Accessed 17 Feb 2024].

Mailgun, n.d. AI and Phishing. Available at: https://www.mailgun.com/blog/email/ai-phishing/ [Accessed 17 Feb 2024].

Arxiv, 2024. An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach. Available at: https://arxiv.org/html/2402.13871v1 [Accessed 17 Feb 2024].

Dark Reading, n.d. AI Cybersecurity: Promise and Challenges. Available at: https://www.darkreading.com/cybersecurity-analytics/ai-cybersecurity-promise-challenges [Accessed 17 Feb 2024].

Microsoft, 2022. Email security with Threat Explorer and Real-time detections in Microsoft 365. Available at: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/threat-explorer-email-security?view=o365-worldwide [Accessed 27 Feb 2024].

Microsoft, 2022. Understanding detection technology in the email entity page of Microsoft 365. Available at: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/step-by-step-guides/understand-detection-technology-in-email-entity?view=o365-worldwide [Accessed 27 Feb 2024].

PR Newswire, 2021. SlashNext Launches Email Spear Phishing Detection and Response for Microsoft 365. Available at: https://www.prnewswire.com/news-releases/slashnext-launches-email-spear-phishing-detection-and-response-for-microsoft-365-301366100.html [Accessed 27 Feb 2024].

Microsoft, 2022. Tune anti-phishing protection. Available at: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/anti-phishing-protection-tuning?view=o365-worldwide [Accessed 27 Feb 2024].

Microsoft, 2022. View email security reports in the Microsoft Defender portal. Available at: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/reports-email-security?view=o365-worldwide [Accessed 28 Feb 2024].

Microsoft, 2022. Anti-spoofing protection in EOP. Available at: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/anti-phishing-protection-spoofing-about?view=o365-worldwide [Accessed 28 Feb 2024].

Reddit user, 2023. Exchange online detecting legitimate email as phishing. Available at: https://www.reddit.com/r/Office365/comments/15eglx4/exchange\_online\_detecting\_legitimate\_email\_as/ [Accessed 2 Mar 2024].

PhishProtection, 2023. Microsoft 365 Being a Soft Target for Scammers- Email Do's and Don'ts. Available at: https://www.phishprotection.com/phishing-awareness/microsoft-365-soft-target-for-scammers-email-dos-and-donts [Accessed 2 Mar 2024].

Google Patents, 2016. Phishing Detection Using NLP and Ontological Analysis. Available at: https://patents.google.com/patent/US20160344770A1/en [Accessed 7 Mar 2024].

Visua, n.d. Anti-Phishing Detection with Visual AI. Available at: https://visua.com/use-case/anti-phishing-detection-with-visual-ai [Accessed 7 Mar 2024].

Egambaram, K.A., Sridevi, R. and Ramesh, D., 2021. Secure and Efficient Authentication Scheme for Access Control in Mobile Pay-TV Systems. Electronics, 12(20), p.4261. Available at: https://www.mdpi.com/2079-9292/12/20/4261 [Accessed 7 Mar 2024].

Muaaz, M. and Mayrhofer, R., 2020. An Analysis of Different Approaches to Gait Recognition Using Cell Phone Based Accelerometers. International Journal of Interactive Mobile Technologies (iJIM), [e-journal] 14(4). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7581503/ [Accessed 7 Mar 2024].

Google Cloud, 2021. Phishing Hunting: How to Identify Threats in Industrial Emails. Available at: https://cloud.google.com/blog/topics/threat-intelligence/phishing-hunting-industrial-emails [Accessed 10 April 2024].

IronScales, 2020. Developing a Machine Learning Model to Identify Phishing Emails. Available at: https://ironscales.com/ironscales-engineering-corner-blog/developing-a-machine-learning-model-to-identify-phishing-emails/ [Accessed 12 April 2024].

ScienceDirect, n.d. Phishing Detection. Available at: https://www.sciencedirect.com/topics/computer-science/phishing-detection [Accessed 10 April 2024].

Uzun, E., Sezer, O.B. and Bicakci, K., 2021. A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques. ResearchGate. Available at: https://www.researchgate.net/publication/361300183\_A\_Systematic\_Literature\_Review\_on\_Phishing\_Email\_Detection\_Using\_Natural\_Language\_Processing\_Techniques [Accessed 10 April 2024].

TechTarget, n.d. GPT-3. Available at: https://www.techtarget.com/searchenterpriseai/definition/GPT-3 [Accessed 16 April 2024].

OpenAI, n.d. GPT-3 Apps. Available at: https://openai.com/blog/gpt-3-apps [Accessed 16 April 2024].

Khurshid, M., n.d. Chat GPT: Exploring Capabilities of AI Language Models. Available at: https://www.linkedin.com/pulse/chat-gpt-exploring-capabilities-ai-language-models-murtaza-khurshid [Accessed 16 April 2024].

Btech, n.d. ChatGPT: A Revolution in Natural Language Processing. Available at: https://btech.id/en/news/chatgpt-a-revolution-in-natural-language-processing/ [Accessed 16 April 2024].

Microsoft Tech Community, n.d. Comparing GPT-3.5 & GPT-4: A Thought Framework on When to Use. Available at: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/comparing-gpt-3-5-amp-gpt-4-a-thought-framework-on-when-to-use/ba-p/4088645 [Accessed 16 April 2024].

ResearchGate, n.d. Exploring the Capabilities and Limitations of GPT and Chat GPT in Natural Language Processing. Available at: https://www.researchgate.net/publication/370048062\_Exploring\_the\_capabilities\_and\_limitations\_of\_GPT\_and\_Chat\_GPT\_in\_natural\_language\_processing [Accessed 16 April 2024].

Appendix)

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated