University of Bedfordshire

Amjed Ashour

1939195

An Investigation into a New Automated Approach to Phishing Awareness

BSc (Hons) Computer Security and Forensics

Undergraduate Thesis Report

Faculty of Creative Arts, Technologies and Science

University of Bedfordshire

Dr Khalid El-Hussein

AY22/23

Amjed Ashour
1939195

Abstract

A Phishing attack is one of the many types of social engineering attacks that are aimed at manipulating users into willingly sharing their sensitive information, such as login details, credit cards information, and other personal data. This information can allow threat actors to have unchallenged access to various confidential accounts belonging to the victims.

Finding a solution to social engineering in general and the various types of phishing, in particular, has evidently proven to be a challenging task, many mitigation solutions have been implemented, but they remain insufficient. One of the reasons is that these solutions usually ignore the cognitive and contextual reasons behind susceptibility to phishing attacks.

In this paper, we investigate the relation between the cognitive processing methods chosen by the victims and the contextual circumstances that, when combined, lead to the victimization of individuals, especially in the workplace.

We also propose a solution that combines the efficacy of security awareness campaigns, and the processing power of computers to efficiently and contextually educate end users on how to detect and avoid phishing attempts, especially in high cognitive pressure situations.

Acknowledgments

The author would like to express their gratitude to their friends and family for their continuous and unwavering support and understanding during this challenging journey.

Key Words: Phishing, Automation, Human Factors, Cybersecurity, Social Engineering, Security Awareness Campaigns.

Contents

List of Figures

List of Tables

Abbreviations

Table 1: List of Abbreviations

Abbreviation	Full Text
IP	Internet Protocol.
URL	Uniform Resource Locator.
SCAM	Suspicion, Cognition, and Automaticity Model .
SP	Signal Probability.
PDF	Portable Document Format.
CTR	Click-Through Rates.
ISAT	Information Security Awareness Training.
CBMT	Contextual-Based Micro-Training.
API	Application Programming Interface.
GSB	Google Safe Browsing.
DOM	Document Object Model.
UML	Unified Modelling Language.
WPM	Words Per Minute.

Introduction

Phishing attacks are the most common and most successful attack vectors that facilitates data breaches with an average cost of 4.91 million dollars per breach and with a total of 19% of overall attacks recorded in 2022. (Mansfield-Devine, 2022)

This paper aims to study the reasons behind the inefficiency of the current deployed solutions and to provide a possible solution that is dynamic and is a combination of both security awareness training and technical-based solutions.

This section provides a background on phishing attacks, and cognitive processing methods used by individuals while reading an email and the relation between the context that an individual is in and the level of their susceptibility to phishing attacks.

Background

Phishing attacks are a form of a social engineering attack aimed at tricking victims into exposing their data to a threat actor either by directly giving the data to the attacker or by a cleverly designed email or message that forces the victim to click a link or download a file that is of a malicious nature that allows the threat actor to gain access to the victim's data or in an organisational context, give the threat actor access to the internal network of the organisation. (Alkhalil et al.,2021)

Over the years, researchers and security practitioners have developed and deployed countless products and solutions to try to limit the success of phishing and spear-phishing attacks. Yet, organisations and individuals are still vulnerable to these types of attacks. The majority of these solutions do not cover the psychological attitude that is present in phishing emails, generally, humans have limited attentional resources while processing information (Steves et al.,2020), this suggests that the presence of a "hidden" cue such as authority or urgency in an email will likely to take over the limited attention that an employee has and force them to rely on their heuristic judgement in processing the email. Therefore, contextual relevance of an email has a huge effect on the ability for an employee to detect a suspicious attempt, an employee who is responsible for data entry will not be affected by a phishing email that is related to payments and it will likely raise suspicions, on the other hand if the same email was sent to an employee who is in the accounting department, it likely will have a higher chance of succeeding.

(Williams et al.,2018) argued that the contextual relevance is the main tool the employees use to interpret the cues present in an email.

Some of the current solutions to combat phishing are purely technical, such as spam filtering, IP and URL blacklisting, and firewalls, as effective as those solutions proved to be, an unforeseen side effect has emerged that is directly connected to the psychological attitude of the employees, (Sawyer & Hancock, 2018) argued that the Human-Computer “team” has caused a Prevalence Paradox in cybersecurity where the users massively rely on computers to protect them from malicious activity, Sawyer argued that the Prevalence effect may alter the user’s strategies and behaviour in ways that will create an attack vector. Spam filters and machine learning malware detection software is successful enough to have the user’s trust, however when the user’s trust becomes ultimate trust, it degrades the user’s ability to detect malicious cues in an email, making it harder and harder to effectively train users on phishing detection and renders campaign awareness a waste of time and money and since machine learning algorithms and spam filters are designed by humans, although they are on the higher end of awareness, are still going to be vulnerable yet trust worthy enough to still cause the prevalence paradox.

As cited previously, Contextual relevance is extremely important in detecting phishing cues, the psychological attitude of the recipient is as important as the contextual relevance and the development of fully automated solution, as alluring as it sounds, is still a contributor to the phishing problem, we believe that the most efficient solution, both technically and financially, is a combination between the human and the computer, so we propose a possible solution by designing a software that will provide real-time evaluation to emails and providing a detailed, yet easy to understand explanation of a potential phishing email and the cues the software used to flag the possible phishing attempt. We believe that this will provide an extra layer of protection that is not only behind the scenes but also actively educating users who might otherwise be either uninterested or simply missing cues while simultaneously building a database of phishing emails that other researchers can use to further study the issue.

Problem Statement

The problem that will be addressed in this research is the lack of consideration of the contextual relevance and the psychological attitude of employees in an organisation in the phishing awareness campaigns and training sessions.

In the context of organisations, one of the current solutions is to train employees on how to safely use the internet by conducting phishing awareness campaigns. These training sessions and campaigns are designed to educate the employees about the risks and consequences of a phishing attack and teach them how to detect a possible phishing attempt.

A phishing campaign is typically structured as follows: first stage, employees are provided with educational materials that explains the premise of phishing, how it works and what falling for it might cause. Second stage, employees go through a period of training sessions, to demonstrate how an employee can recognise a phishing attempt by educating them on what a suspicious link or attachment looks like. Third stage, an unannounced simulated phishing attack, designed to test the efficiency of the first and second stages and gauge their benefits, a phishing email is sent to a random sample of the trained group to assess their abilities to detect or recognise a suspicious email and based on the results, a retraining or an organisation-wide email might be sent as a reminder to the employees on how to recognise a phishing attempt and the available reporting procedures available in the organisation.

In theory, this solution should work, in a controlled environment such as a university or a lab setting this will work, yet phishing attacks still succeed and are still causing tremendous amounts of harm to organisations worldwide. Figure 1 shows the rise in the number of phishing attacks during the years 2020, 2021, and 2022 (the pandemic duration)

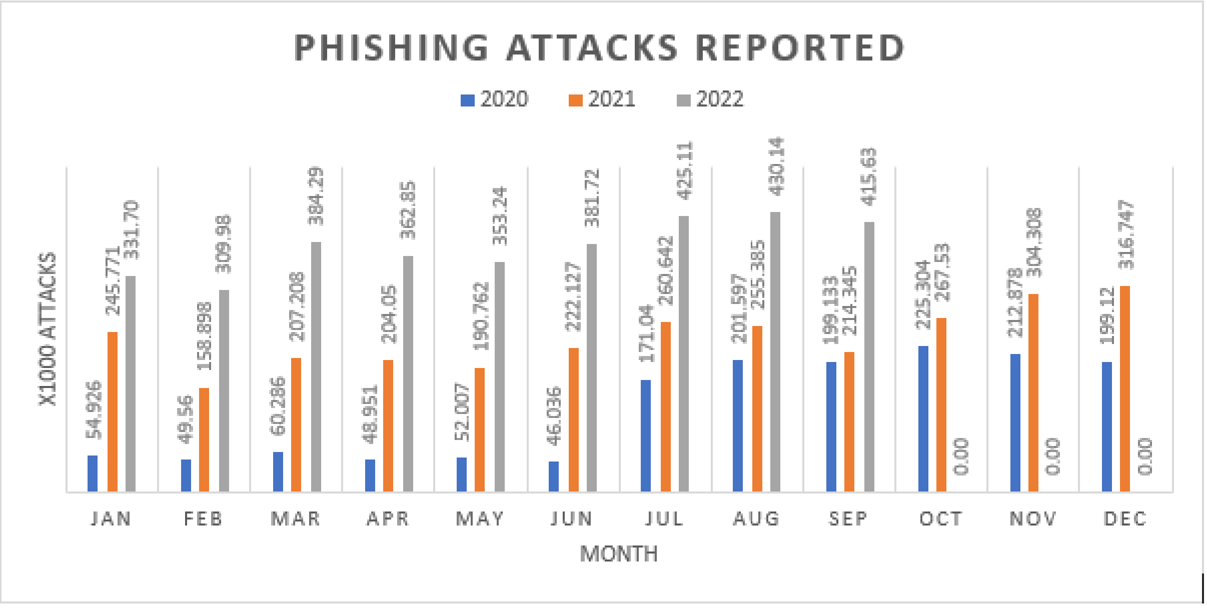


Figure 1: A Monthly comparison between the number of phishing attacks in the period between 2020 - 2022 (APWG, 2020-2022)

Figure 2 shows the average cost of a data breach in the years 2021 and 2022.

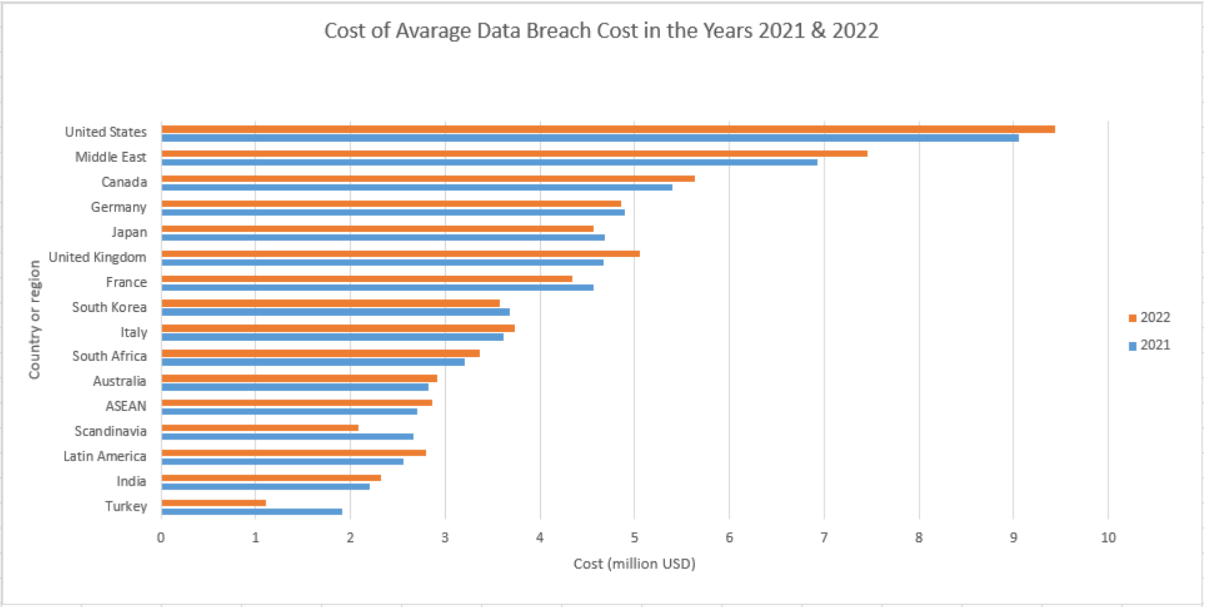


Figure 2: Average Cost of A Data Breach for The Top 17 Countries and Regions in 2021 & 2022. (Mansfield-Devine, 2022)

Scope and Limitations

1. Scope

The scope of this artefact is to develop a software in the form of a web browser extension that scans received emails and analyse them to detect phishing attempts based on the intent of the text in the body of the email, a scan for the links and attachments in the email and the source of the email by using multiple APIs then flag the email to be either legitimate or malicious based on the previously mentioned criteria.

2. Limitations

Time constraints

Due to the limited amount of time, some compromises were made to ensure the delivery of a fully functional software, some functions, such as the extension control that was aimed to allow IT administrators to tailor the extension based on the policies implemented in the organisation was removed from the scope.

Testing and Evaluation

The full extent of the impact this software might have on phishing awareness might need years to be clear, the testing process that will be implemented is going to test the accuracy and speed of the responses generated by the software.

User acceptance

In general, the success or failure of a software is tightly tied with users accepting it, to be able to accurately test the impact of the software will highly depend on the software being adopted by an organisation to measure the impact of it on their employees.

Data

Acquiring data about the effects of a phishing campaign and training sessions to compare them with the results of this artefact is difficult, this kind of data is usually classified and is only available to a limited number of professionals in an organisation.

Literature Review

Introduction

In their process of initiating an attack, threat actors analyze the target's attack surface and divide it into a number of attack vectors in an effort to increase the likelihood of a successful attack. Emails, a common attack vector across the majority -if not all- organisations, is often exploited by threat actors(Singleton, 2022).

Organisations and security researchers have been developing and implementing countermeasures against these threats, but the vulnerability remains. This is partly due to the static nature of the solutions or training programs that have been developed where they teach employees about the risks associated with phishing and spear phishing and how to identify them. however, despite these efforts, phishing is still an issue which suggests that there is a flaw in the design of such training and awareness campaigns.

Designing an effective training and awareness campaign that accounts for all of the factors that affects susceptibility to phishing is an incredibly challenging task. The goal of this literature review is to highlight these factors and how they contribute to the limitations of current solutions, in order to propose a potential solution that addresses these gaps.

1. The SCAM

Vishwanath et al. investigated the psychological factors that might affect an individual's susceptibility to phishing attacks, in their paper ***Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility*** they aimed to explore how these three factors (Suspicion, Cognition, and Automaticity) affected a user's interaction with an email.

Suspicion in this context is defined as the level of trust an individual has towards the email, Cognition is the process of analysing the email and the ability to detect malicious intent in the email, and Automaticity is the reliance on a heuristic or a habitual interaction with the email (Vishwanath et al., 2016). Vishwanath argued that these factors are the main psychological factors that affect an individual's ability to detect phishing attempts.

The SCAM (Suspicion, Cognition, and Automaticity Model) proposes that individuals with high levels of suspicion, and cognition abilities are less susceptible to a phishing attack, as well as reducing the automaticity while responding or interacting with an email. Vishwanath proposed 4 hypothesises, A higher level of heuristic processing is likely to decrease the suspicion, A higher level of systematic processing is likely to increase suspicion, Cyber-risk beliefs will be negatively related to heuristic processing, and cyber-risk beliefs will be positively related to systematic processing. To test the hypotheses, Vishwanath et al. conducted two experiments of the same nature, where 2 separate university classes in two different semesters (Spring, and Summer) were recruited and given email addresses, the two groups received an identical email with the only difference is the method of malware delivery (first group was a hyperlink, and the second group was a PDF attachment). The results confirm the hypotheses, individuals who had a higher level of suspicion were less likely to fall victims, and the processing system the individual's used, highly affected the outcome, where if the individual believed that their actions were risky, they used a systematic approach to analysing the email, while individuals who believed that their cyber actions were safe, used a heuristic approach. Furthermore, individuals who relied on their habitual patterns while processing the email showed a significant reduction in their abilities to detect deception, leading to individuals missing cues that could potentially arouse suspicion.

The study proves that external factors (i.e., Email habits, and cybersecurity knowledge) have a considerable effect on the susceptibility to phishing, factors that cannot be controlled nor predicted efficiently in a phishing awareness campaign.

1.1. Processing

Previous research identified two main cognitive methods of processing, first, Systematic processing, a method where individuals examine information carefully, looking through the facts presented in the information in detail and arriving at a conclusion based on what they perceived from analyzing the information. Second, Heuristic processing method, where individuals rely on their experience or what they already know to arrive to a conclusion.

The problem individuals face is that reading an email has become a habit rather than a purposeful task. Tasks that are continually enacted in a stable environment and conditions turn into routines, and over time, these routines become habits, an automatic function that is applied whenever the task is performed, leading to the use of heuristic processing. Heuristic processing, although considered efficient and economic, processing resources wise, was observed to lead to an increase in judgment errors and irrational thinking.

On the other hand, systematic processing requires a lot of cognitive resources when used while processing information, but the results are much more rational and less likely to cause errors in judgment.

In an email that has cues of deception, it is important to use the systematic processing method. The cursory skimming or reading of an email that is associated with heuristic processing is done without conscious, and without regard to the consequences since it does not require a large amount of cognitive processing resources.

1.2. Suspicion

Suspicion is defined as the act or instance of suspecting something is wrong without proof or on slight evidence: MISTRUST (Merriam-Webster, 2023).

Vishwanath (2016) argued that suspicion is a necessity for detecting deception, when aroused, even a small amount of suspicion was observed to improve the deception-detection abilities that a user possesses, making it fundamental to the process of deception-detection.

The effects of phishing cues in an email varies, viewing a certain cue as suspicion inducing or compelling, highly relies on the individuals themselves and on the context the individuals are in when processing an email (Steves et al., 2020), deception cues in a highly contextually relevant email proved to be extremely difficult to detect, the more contextual relevance, the less likely the cues will arouse suspicion.

Previous literature around the psychological reasons behind falling victim to phishing attacks concluded that poor cognitive processing is the key factor, the majority of solutions focused on educating individuals on how to effectively detect deception, however, the habitual behaviour of individuals outweighed the effect of the training sessions in a relatively short period of time. The model proposed by Vishwanath, accounts for the different methods of cognitive processing, and the reasons as to why individuals, or in this case employees chose to use a certain processing method.

2. The Prevalence Paradox

The prevalence paradox in cybersecurity is a term used to describe the effects of the high levels of success that automated solutions that are aimed at protecting humans from malicious or unwanted emails (i.e., spam email filters) on the human's ability to detect malicious emails due to the rarity of encountering them (Sawyer & Hancock, 2018). The authors argued that the increasing success of automated solutions has led to the degradation of humans' ability to detect the malicious cues present in the emails they daily encounter.

(Sawyer & Hancock, 2018) argued that these effects are more than just a possible nuisance, but a potential attack vector that can be utilised by threat actors. (Sawyer & Hancock, 2018) referred to the cues in the email as Signal Probability (SP), the authors argued that in an environment where the Signal Probability (SP) is low, the prevalence paradox effect is at its highest, since the user is ultimately expecting the automated system to protect them.

(Sawyer & Hancock, 2018) tested their hypothesis by creating a simulated fictitious organisation, Cog Industries. 30 participants assumed the roles of administrators in the company who would receive emails either containing or requesting PDF files containing sensitive information, the participants had three options, download the PDF attachments, upload PDF attachments, or report malicious activity. The participants attended a brief training session containing 20 emails, either legitimate or malicious emails, and were required to pass a test with 80% or more, all participants passed the test.

The main experiment had a sample of 300 emails that the participants were instructed to choose from and to take their time doing so. The results confirmed what (Sawyer & Hancock, 2018) hypothesised, regardless of the type of attack they encountered, participants who picked emails with low Signal Probability (SP) scored the lowest accuracy and fell victims of a malicious attack.

The main outtake from this study is to highlight the fact that threat actors may use this human cognitive vulnerability rather than attacking a machine. As shown, this is certainly an easier alternative for the attacker to follow.

As automation of defences become more prevalent and effective, the Prevalence Paradox can potentially become amplified. arguably, current solutions are overlooking the human' vulnerabilities in favour of technical solutions, which widens the gap between the humans and the machines, two variables that are the main targets of a threat actor. Focusing on further developing the automated solutions alone is only increasing the imbalance of the focus and attention dedicated to technical solutions. In a limitless resource environment, this will not be an issue, but in reality, this is causing a disconnect between the human and the machine in the context of preventing phishing attacks.

The increase in the success of automated security solutions resulted in an unforeseen, and unwanted side effect. The increased reliance on automated solutions to protect against threat actors resulted in the degradation of the individual's ability to detect deception, the fact that individuals are almost fully trusting that the technical solutions are perfect or are going to "catch" all incoming threats has rendered individuals blind to non-obvious deception cues, the more scares the cues are, the more likely to fall victim to an attack.

3. Susceptibility in the Workplace

Williams et al., (2018) conducted two studies to try and determine the reasons of susceptibility to phishing attacks in the workplace and the effects other factors might have on it. The studies showed that the presence of authority and urgency cues in an email increased the likelihood of falling victim to such attacks.

Study one aimed to examine how the presence of authority and urgency cues affected the recipients of an email. Authority cues work on mimicking individuals who have authority in an organisation, and urgency cues are an indication of some sort of a time constraint that is tied to a certain task in the email.

In this study, Williams suggested two hypotheses, first, that the presence of urgency cues within a simulated spear phishing email is more likely to increase the susceptibility to phishing in the workplace. And second, the presence of authority cues within a simulated spear phishing email will likely increase the susceptibility to phishing.

Williams used a dataset of a phishing simulation from a UK public sector organisation with more than fifty thousand employees that deals with members of the public and handles their sensitive information. The dataset was collected based on nine simulated phishing emails that were sent to all employees within the organisation.

The emails were sent from a fictional organisation and were designed to mimic actual phishing emails previously received by employees in the organisation, the emails were structured as follows, first, the emails included the name of the recipient (i.e., 'Dear John'), all emails contained the fictional company's logo, and all emails contained a link that the recipients were encouraged to click.

The results of study one confirmed both the hypothesis suggested by Williams, the presence of authority and urgency cues increased the susceptibility to phishing, however the study showed that the presence of authority cues increased the likelihood of clicking the link by approximately 100%.

Study two was designed to cover the missing factors that were not provided by the organisation in the original dataset such as the situational-level and individual factors, and to examine whether external factors that are not present in the email itself had any effect on the susceptibility to phishing.

Study two aimed to explore factors that are external to the email itself (i.e., the contextual relevance of the job description). The study used a different approach than the one used in study one, instead of using datasets, Williams used a focus group methodology to determine the effect of the external factors on the susceptibility to phishing and spear phishing. The results showed that the external factors play a very important role when it comes to the level of susceptibility to phishing. The degree of knowledge possessed by the participants, the contextual relevance of the job description, and the work routine all affected participants' ability to detect a malicious email, both in a negative and positive way.

In general, Williams' research showed that susceptibility to phishing emails is not limited to the contents of the email, rather the opposite, the external factors that are related to the recipient are a bigger factor and are a key component to be considered while studying the shortcomings of the current solutions to social engineering in general and in phishing and spear phishing in particular.

Despite the use of awareness campaigns and trainings, employees are still vulnerable to phishing attacks, specifically because of the lack of consideration for the external factors in these campaigns. The email itself, the job description, the context is which the email was received, and who the recipient is, are all important factors that affect the susceptibility to phishing attempts and are extremely important factors that affect the cognitive processing method employees chose to use while processing the information in an email.

4. The Effects of Awareness Campaigns

Carella A. (2017) studied the impact of security awareness training on the click-through rates (CTR) within the phishing attack vector. Carella A., measured the impact by testing three separate groups of students where each group had a different approach to the problem.

Carella A., concluded that the awareness training and campaigns do in fact have a considerable and important effect on the CRTs.

Carella A. separated 150 students into three groups, a control group (A) where the students get no training at all, a group (B) where the students will receive the training on a weekly basis via a documents explaining the anti-phishing techniques with each phishing email, and a group (C) where an in-class 30 minutes presentation delivered by one of the researchers will explain the importance of secure internet usage in depth.

The control group (group A) received four separate phishing email over the course of the study and their results will be used to compare, first the effects of security awareness training, and second to compare the different methods of delivering the training,

The documentation group received the documents persistently with each phishing email sent, if a student clicked the link in the email, they were redirected to a website that contained an awareness message informing them that they have fallen to a phishing attempt in order to encourage them to read the document. And lastly the in-class presentation (group C), was designed to teach the students about the safe usage of the internet and its importance with a focus on phishing and anti-phishing techniques as well as an in-class exercise where the students were guided through it by the presenter.

The study was conducted over a 7-week period, where one phishing email sent per week, the results were as follows, group (A) scored the lowest in the entire duration with the highest CTR throughout with an average percentage of 52% CTR, group (B) scored the highest over the course of the study with a 26% CTR, and group (C) scored in the middle with 46% rate.

On a week-by-week basis, group (A) had a consistent CTR with a around the 52% with a plus or minus 2% variation, group (B) had a considerable drop in the CTR where the first week they scored a 50% and last week 8% with an average drop of 6% per week, and group (C) showed a decrease in the CTR over the first 3 weeks and then an increase to the original CTR over the last 4 weeks, however, group (C) scored the highest decrease in CTR by 16% in the first week.

The approaches used showed interesting results, persistent training showed persistent improvement throughout the duration of the study, the majority of students who fell victim to the first phishing email sent, did not fall for the second one, that is true for all phishing emails over the course of the study. Interestingly, students who received in-class training showed the highest rate of improvement between the first and the second phishing emails, further investigation needs to be conducted in order to determine the impact of in-class training to gauge its impact if it was on a week-to-week basis.

In conclusion, the study shows that security awareness training campaigns, although they do not mitigate the issue completely, have a considerable effect on susceptibility to phishing. It shows the different effects different approaches have on the CTR and highlights the importance of the security training in general.

5. Evaluation of Contextual-Based Training

In their study Evaluation of Contextual and Game-Based Training for Phishing Detection, Kävrestad (2022) aimed to evaluate the efficiency of the methods used to deliver Information Security Awareness Training (ISAT), the methods Kävrestad chose to evaluate are Game-Based training, where participants learn through playing a game, and Context-Based Micro-Training, where users were given contextually relevant instructions on how to deal with a phishing email at the time of receiving the phishing email.

The study consisted of 41 participants separated over three groups, control group, Game-Based group, and Contextual-Based Micro-Training group (CBMT), all three groups were given access to an inbox and were asked to classify the emails within it into malicious or not and were scored based on how accurately they classified the emails.

The participants in the Game-Based group played an educational game developed by Google, the game covered all the identifiers that might be present in a phishing email. The CBMT group participants received a training developed by the research team, written information that is presented in a prompt upon opening the inbox, which included a number of general tips on how to avoid falling victim to a phishing attack and encouraged the users to take part in extra training by following a link that led to a text-based presentation.

The results show that the participants in the CBMT groups scored higher than the other two groups, with 21.4% of the participants in the CBMT group got a perfect score where on the other hand, no participants in the other two groups got a perfect score. The overall statistical analysis of the results also showed that the CBMT group scored higher than the other groups in general.

Kävrestad suggests that the reason for the CBMT group success is attributed to the fact that the CBMT provides mechanisms that increase awareness in addition to the training itself while the Game-Based training only provides training materials.

To summarise, the study evaluated two methods of delivering anti-phishing training, Game-Based training method, and a Contextual-Based Micro-Training method. The study concluded that between both methods, the CBMT was more effective in improving the individuals' accuracy in detecting phishing emails or malicious intent, however, the study also suggested that even with the training, the percentage of individuals who are able to correctly identify phishing emails is relatively small, assuming that the criteria for avoiding falling victim to a phishing email is to correctly identify 100% of the phishing cues or identifiers present in the email.

6. Analysis

For years, researchers have attempted to identify the exact reasons an individual is susceptible to phishing attacks in order to develop a "once and for all" solution, theoretical frameworks and models, such as the Phish Scale (Categorizing human phishing difficulty: a Phish Scale: Steves M. et al., 2020), and the SCAM (Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility: Vishwanath et al., 2016), these types of research all point towards the recognition of the role of cognitive behaviours and the relationship between them and susceptibility to phishing. These papers explored the reasons beyond the contents of an email, they studied the socio-psychological reasons that are common between individuals who were, and are, being victimised by phishing and spear phishing attacks.

What these studies have identified is a number of common behaviours that are observed in the majority of the victims, the cognitive processing methods that they use while performing a repetitive task such as reading an email, most of the victims used a heuristic processing method that is associated with processing habitual tasks. Williams et al., (2020) suggested that the issue of the heuristic processing can be extended beyond an individuals' issue, where threat actors can use it as an attack vector by designing emails that contained cues that triggered that method of processing such as authority, and urgency cues.

Williams et al., also studied the effects that the current security awareness campaigns have on employees, the results were conflicted with the goals of these campaigns, some employees admitted that they are overwhelmed with the amount of information they receive from these campaigns, and the reason being that in an organisation, there might be multiple different campaigns running at the same time, such as health and safety campaigns, or a daily disturbance that will take a large amount of the cognitive processing from the already limited processing resources an employee have.

On the other hand, Kävrestad et al., in their study ***Evaluation of Contextual and Game-Based Training for Phishing Detection*** (2022), concluded that such campaigns are the most effective when it comes to educating individuals on how to detect deception cues in an email with consistent results throughout the course of the study.

Kävrestad et al., (2022) also found that the quickest way to achieve efficient deception detection techniques was by introducing a Contextual-Based Micro-Training (CBMT), however, the results showed that in order to achieve consistent success with the CBMT, it must be implemented on frequent basis, which demands the dedication of a significant number of resources.

Kävrestad et al., findings are of significant importance, there is no doubt that security awareness campaigns do have a positive impact on susceptibility to phishing, Carella A. et al., proved that in their study ***Impact of Security Awareness Training on Phishing Click-Through Rates*** (2022), the results of the study showed a huge disparity of click-through rates between trained and untrained individuals regardless of the training method used.

Some other solutions that were proposed to combat phishing, are the fully automated solutions, where the objective of detecting deception or malicious intent in an email is solely on the shoulders of a software deployed either on an email server or the end user's machine, unfortunately, this approach does not achieve the desired goal, evident by the increase in successful phishing attacks over the years (see Figure 1), not only that, it might also have an unwanted side effect where the individuals rely heavily on the automated solution that their ability to detect deception is degraded to a point where they become blind to phishing cues that do make it through past the automated solutions (Sawyer & Hancock, 2018).

As far as we are aware, there are no solutions that combine both the automated solutions and the CBMTs, which is what this artefact is aimed at, by combining both, we hope that we can reach middle grounds between alleviating the taxing tasks on cognitive resources and educating end-users simultaneously.

Methodology

Introduction

In this section, we will discuss the design aspects of the artefact and the reasoning behind the decisions made, providing a comprehensive overview of how this artefact aims to offer a possible solution to the phishing susceptibility in a work environment.

Design

During the design stage, we followed an iterative process which involved an evaluation of each software element and its importance and effect on the outcome, we studied the effects on speed, accuracy, and convenience. The decision landed on using a web browser extension to facilitate the ease of access to both the end-users, and the IT teams in the organisations.

For the backend, we decided to use Python as the programming language as it offers a simpler interaction with the APIs we used and offers a wide range of libraries that facilitated the management of the APIs calls such as Flask to handle the POST requests from the frontend and towards the APIs endpoints.

We decided to use the VirusTotal API, Google Safe Browsing API, and OpenAI API.

VirusTotal API

The VirusTotal API was used to obtain information about the safety of files and URLs that are included in the email body, we used the Python virustotal-api library to handle the request to two end points, URL and File “scan” endpoints, and URL and File “report” endpoints. We filtered out the results that returned a “negative” detection value (No threats found).

Google Safe Browsing API

The Google Safe Browsing API (GSB) was used to validate the results of the VirusTotal API, we used the “threatMatches:find” endpoint to eliminate any false positives that might be present in the response from the VirusTotal “report” endpoints. Any result that came back negative from the GSB was discarded and the result from the VirusTotal API was flagged as a false positive.

OpenAI API

OpenAI API was used to generate text based on the results of both the VirusTotal API and the GSB API, we opted to use the text-davinci-003 natural language model and the “Completion” endpoint to generate text to explain both the negative and the positive results to present the end-user with a natural-sounding prompt.

We refined the generated text using the following parameters:

Table 2: Parameters used in the OpenAI API “Completion” Endpoint

Parameter	Description	Value
Prompt	The text sent to the “Completion” end point	--
Temperature	Controls the creativity of the outputted text	0.7
Max_tokens	The maximum number of tokens to generate in the outputted text	512
N	The number of completions to generate for each outputted text	1
Stop	The suffix that the model will stop generating text at	None
Echo	A Boolean value to indicate if the input prompt is to be used in the output	False
Best_of	Chose the best output from the generated text	3
Frequency_penalty	Controls the degree of token repetitions in the generated	0

	text (higher values means less repetitions)	
Presence_penalty	Controls the frequency at which the model repeats itself (higher values means less repetitions)	0

Prompt Design:

We elected to use two prompts to be used in the development of the artefact, first, a natural language analysis prompt, aimed at analysing the grammar, and spelling errors in the email, as well as any inconsistencies in the text itself or between the different text elements such as an inconsistency between the sender of the email and the signature, or the title of the email and requests in the email body.

And the second prompt is used to analyse the generated text and classify the sentiment of it, two options were specified, either “positive” or “negative”.

For the frontend we decided to use vanilla JavaScript, since it is the simplest way to interact with the DOM elements whether the interaction was reading the DOM, extracting elements from the DOM, or adding elements to the DOM.

UML Diagrams

In this section, we will provide a visual representation of how the software component of the artefact behaves and operate using the Unified Modelling Language (UML). UML is a standardised language used to visually present the design of software solutions.

Use Case Diagram

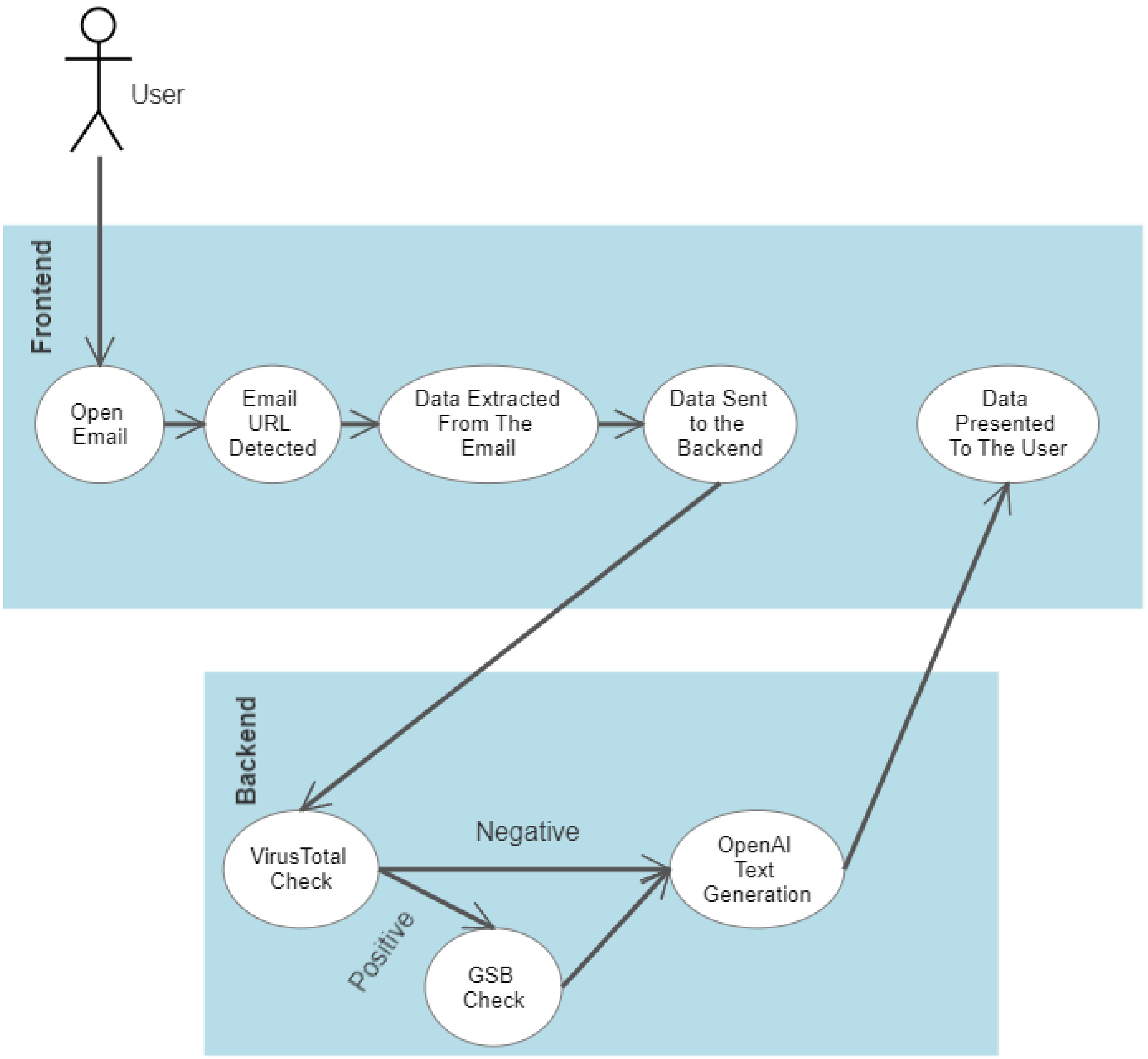


Figure 3: Use Case Diagram

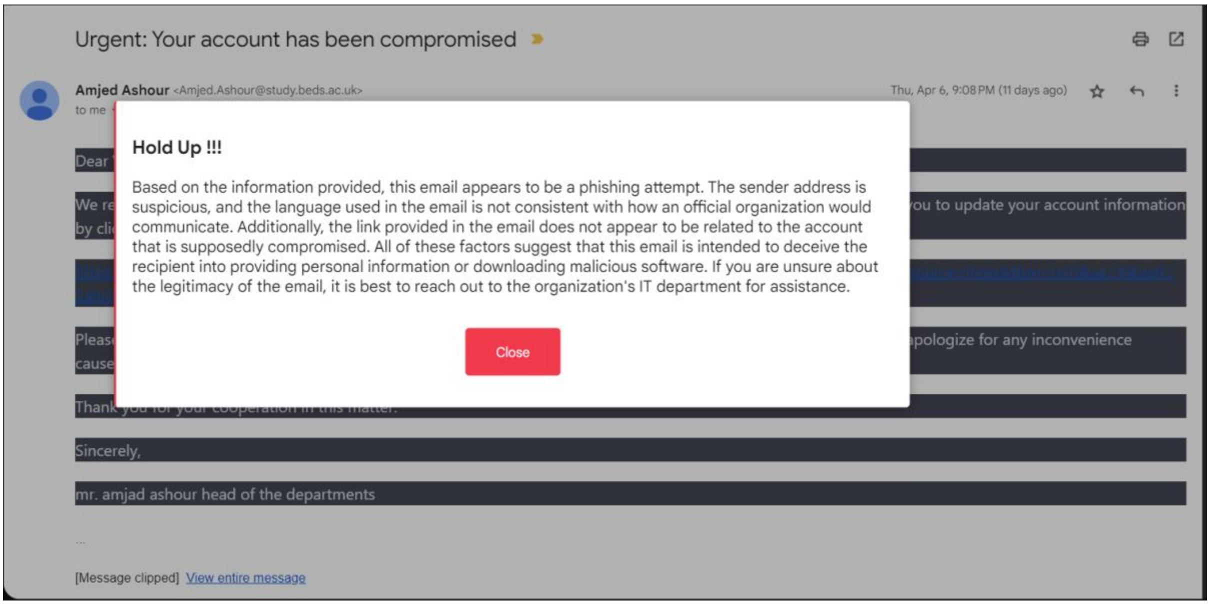
This Use Case Diagram shows the interactions between the different functions in the software, after the initial interaction from the user (Open Email), the entire process is automated.

The browser extension will determine when to start the script based on a regular expression (regex) that is compared with the current URL . When the script is triggered, it will access the DOM elements in the browser to extract the sender, title, main body, hyperlinks, and attachments from the email then it will send a POST request to the backend where it sends the extracted data in a JSON format.

After the data is received by the backend, the VirusTotal API will be used to send a POST request to the “scan” endpoints and it will await a response of a successful scan, then it will send a GET request to the “report” endpoint using the scan ID.

Depending on the results from the VirusTotal API, there are two paths the code can follow, first, if the results received from the VirusTotal API contains a positive value in regards to any of the scanned URLs or attachments, a POST request will be sent to the GSB threatMatches:find endpoint using the GSB API to eliminate the false positives that might be present in the VirusTotal response, if the results from the GSB also come back positive, the email will be flagged as a possible phishing attempt and the results will be sent to the OpenAI API to generate a text that explains the results.

On the other hand, if the VirusTotal response came back negative, the GSB check will be skipped, and the results will be sent directly to the OpenAI API to generate a text explaining the results. The results will then be sent back to the frontend to be presented in one of two ways to the end-user, first, if the results came back positive, a prompt covering the email body will be displayed explaining the results, what the issue is, how it was detected, and some advice to contact the IT department in the organisation (Figure 2).



Urgent: Your account has been compromised

Amjed Ashour <Amjed.Ashour@study.beds.ac.uk>
to me

Thu, Apr 6, 9:08 PM (11 days ago)

Dear

We re ... ou to update your account information
by cli

Pleas ... pologize for any inconvenience
cause

Hold Up !!!

Based on the information provided, this email appears to be a phishing attempt. The sender address is suspicious, and the language used in the email is not consistent with how an official organization would communicate. Additionally, the link provided in the email does not appear to be related to the account that is supposedly compromised. All of these factors suggest that this email is intended to deceive the recipient into providing personal information or downloading malicious software. If you are unsure about the legitimacy of the email, it is best to reach out to the organization's IT department for assistance.

Thank

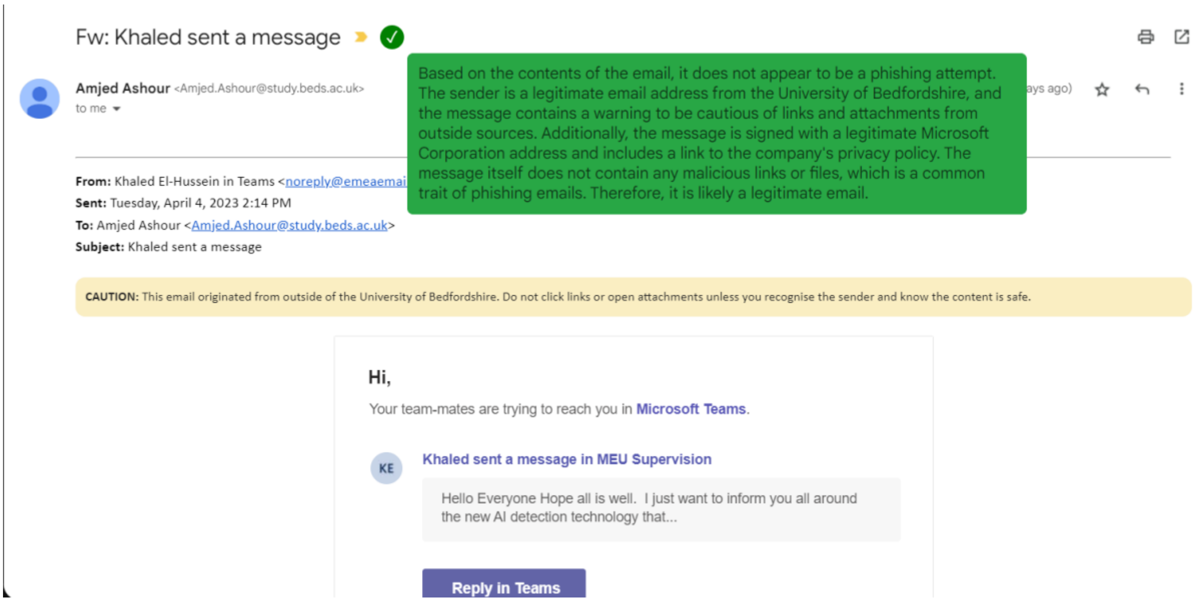
Sincerely,

mr. amjad ashour head of the departments

[Message clipped] View entire message

Figure 4: Example of the Prompt in the case of a possible phishing attempt

The second case is when the results are negative, a green check mark will be displayed next to the title of the email, when hovered over, it will show a tooltip informing the end-user why was the email deemed safe. (Figure 4)



Fw: Khaled sent a message ✓

Amjed Ashour <Amjed.Ashour@study.beds.ac.uk>
to me

Based on the contents of the email, it does not appear to be a phishing attempt. The sender is a legitimate email address from the University of Bedfordshire, and the message contains a warning to be cautious of links and attachments from outside sources. Additionally, the message is signed with a legitimate Microsoft Corporation address and includes a link to the company's privacy policy. The message itself does not contain any malicious links or files, which is a common trait of phishing emails. Therefore, it is likely a legitimate email.

From: Khaled El-Hussein in Teams <noreply@emeaemai
Sent: Tuesday, April 4, 2023 2:14 PM
To: Amjed Ashour <Amjed.Ashour@study.beds.ac.uk>
Subject: Khaled sent a message

CAUTION: This email originated from outside of the University of Bedfordshire. Do not click links or open attachments unless you recognise the sender and know the content is safe.

Hi,

Your team-mates are trying to reach you in **Microsoft Teams**.

Khaled sent a message in MEU Supervision

Hello Everyone Hope all is well. I just want to inform you all around the new AI detection technology that...

Figure 5: Example of the Prompt in the case of a safe email

Modal Design

Functional Design

In order to limit the contribution to the Prevalence Paradox, the popup modal was carefully designed to limit the need for cognitive processing resources, the colours that were used, the method that the information was structured by and the prompt that was used to generate the information were all factors that we considered while designing the modal.

According to (Norman, 2021) in his book ***The Design of Everyday Things***, there are multiple factors that contribute to the understanding of presented information in products, in our case, the modal, for instance, Affordance defines the properties of an object that suggest how it works to intuitively and easily work with the object without the need for extra information or guidance, and that's why the modal allows the user one interaction, which is close the modal. We disabled the Close button for 10 seconds to avoid accidental dismissal of the modal.

Furthermore, Constraints, defined in this context as limiting the user's ability to misuse the product, in this case the modal, (Norman, 2021) argued that a design that does not include constraints is one of the reasons why warnings and instructions fail.

Visual Design

In order to ensure an effective teaching strategy that is both, easy to remember, and requires minimal cognitive processing resources, the modal design consists of one main colour, red, to grab attention, and the text is displayed in black on a white background.

According to a study published on the National Library of Medicine (Dzulkifli & Mustafar, 2013), The black text on a white background proved to be the most efficient when it comes to retention, which is one of the goals of this artefact, in the same study, (Dzulkifli & Mustafar, 2013) found the colours such as red, orange, and yellow (warm colours) have a better effect on attention compared to cool or cold colours.

Prompt Design

In order to produce a clear and concise informative text that can be accessible by everyone regardless of their technical background, the prompt that is used to generate the responses from the OpenAI API needed to be tweaked on multiple occasions, though trial and error, the final prompt is as follows:

"Please analyze the following email and provide your opinion on whether it appears to be a phishing email or not.

the email contains the following information:

Sender: {sender}

Title: {title}

Body: {mainBody}

Based on the contents of the email, please provide your opinion on whether it appears to be a phishing attempt or not.

If you believe the email may be a phishing attempt. Phishing emails often contain malicious links or files that can harm your computer or steal your personal information.

If you're unsure about the legitimacy of an email, you can always reach out to your organization's IT department for assistance.

Please provide an explanation of why you believe the email is or is not a phishing email."

This prompt resulted in the most natural-sounding response, according to (McNamara et al., 1996) an easy to read, coherent and natural text is easier to comprehend and recall, which again is the goal of the artefact.

Results

As discussed earlier, the aim of this artefact is to educate end-users about phishing without using extra cognitive processing resources by combining the information that is included in a typical security awareness campaign and the automatic phishing detection techniques.

Testing Setup

To accurately test the software's capabilities, the test setup included a total of 12 emails divided into 4 categories, *Easy to detect*, *Medium difficulty*, *Hard to detect*, and *Legitimate emails,* Easy to detect emails featured obvious phishing cues such as grammar and spelling errors, inconsistencies between the title and the email body, and misspelled URLs, as well as informal language, the hard to detect category included subtle cues that are not easily detected by human, the structure of the emails mimicked legitimate emails found in the inbox if the author with the addition of imbedded links (i.e., "click here") where the link is not immediately visible for the user to evaluate, and personalised emails that addressed the user by their name instead of "Dear User,", and the medium difficulty category is a combination between the Easy to detect and Hard to detect categories, where the email included both obvious and subtle cues, and lastly, the Legitimate emails category consisted of legitimate emails similar to emails found in the authors inbox, where all real company names and URLs were changed into a fictional company name "aiforums" and a website URL https://www.aiforums.com/, both of which do not exist. The Legitimate category was used as a control category.

Accuracy

The software was able to successfully detect all emails in the Easy to detect, Hard to detect, and Legitimate emails categories, on the other hand, the software's performance in regard to accuracy in the Medium difficulty category was unsatisfactory, with a 33.3% accuracy score.

Results are represented in Figure 5

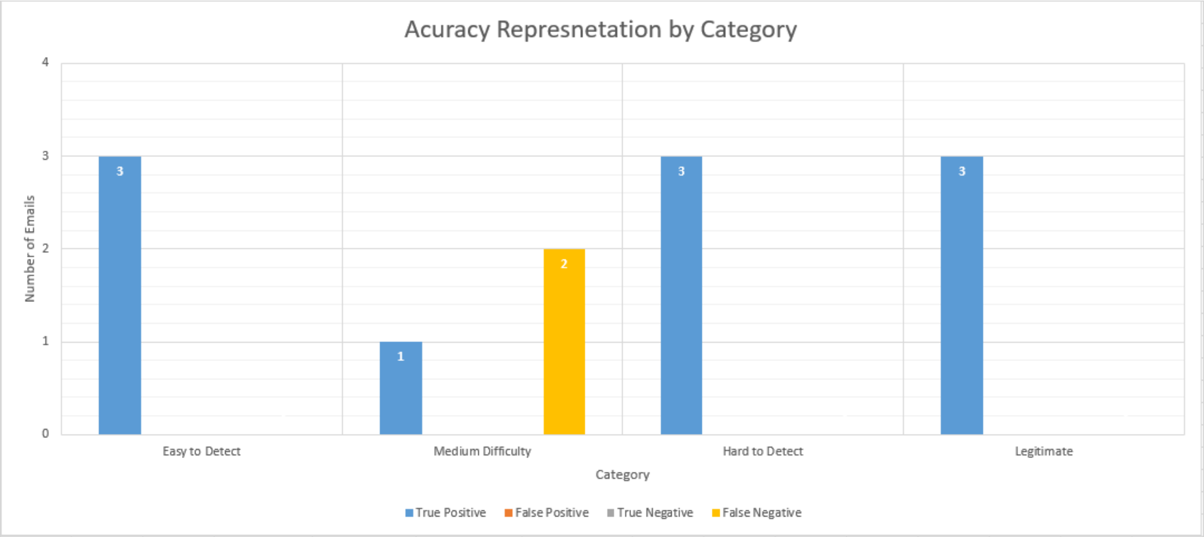


Figure 6: Accuracy Representation by Category.

Table 3: Accuracy Representation by Category

Category	Total No. of Emails	True Positive	False Positive	True Negative	False Negative
Easy to Detect	3	3	0	0	0
Medium Difficulty	3	1	0	0	2
Hard to Detect	3	3	0	0	0

The accuracy of overall detection scored 83.3%, with all False Negatives in the Medium difficulty category which individually scored 33.3% accuracy.

Response Times

Response time is a very important factor to be considered in the fight against phishing, the time needed to open an email and act on it is a very crucial time. If the response from the software was too slow, there is a higher chance that the user would already have clicked on a link or downloaded an attachment. According to (Carver, 1991), the "skimming" reading rate in Typical collage reading rate is approximately 450 Words Per Minute (WPM). As discussed earlier, skim reading, or "skimming" is tied with the heuristic cognitive processing.

Assuming an average word count in a typical email is between 100-200 words, this gives the software approximately 13 seconds to analyse and present a response (for a 100 words email), $Time = \frac{60 \text{ Seconds x } 100 \text{ words}}{450 \text{ WPM}}$ = 13.3 seconds.

We have tested the time needed to present a result for each email in the four categories and the results are represented in Figures 6 and 7

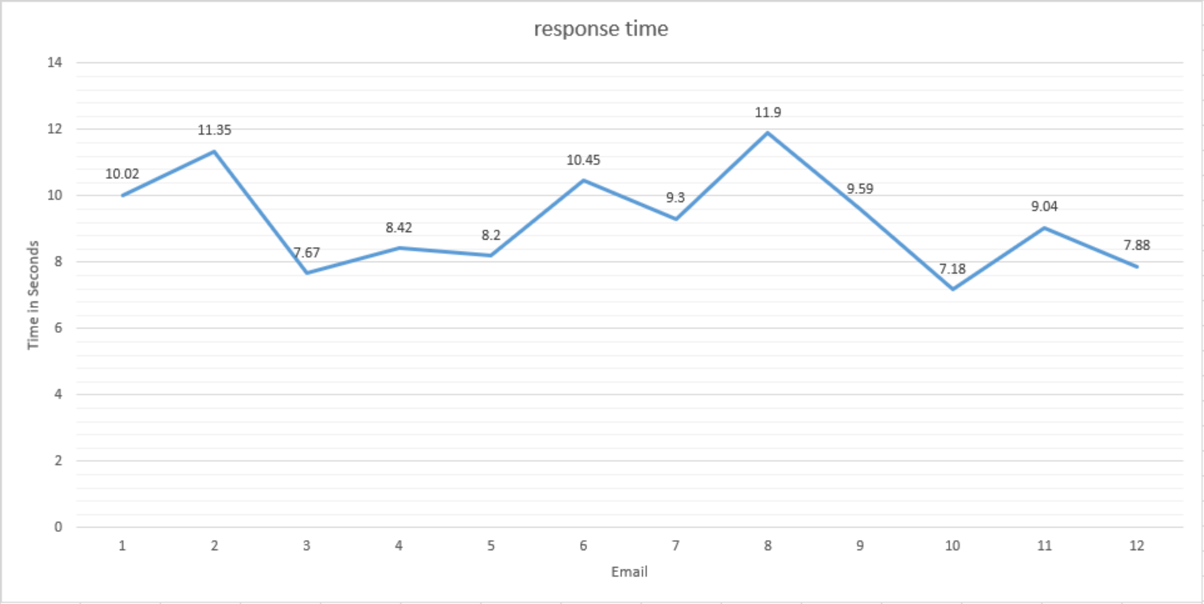


Figure 7: Response Time for the Individual Emails

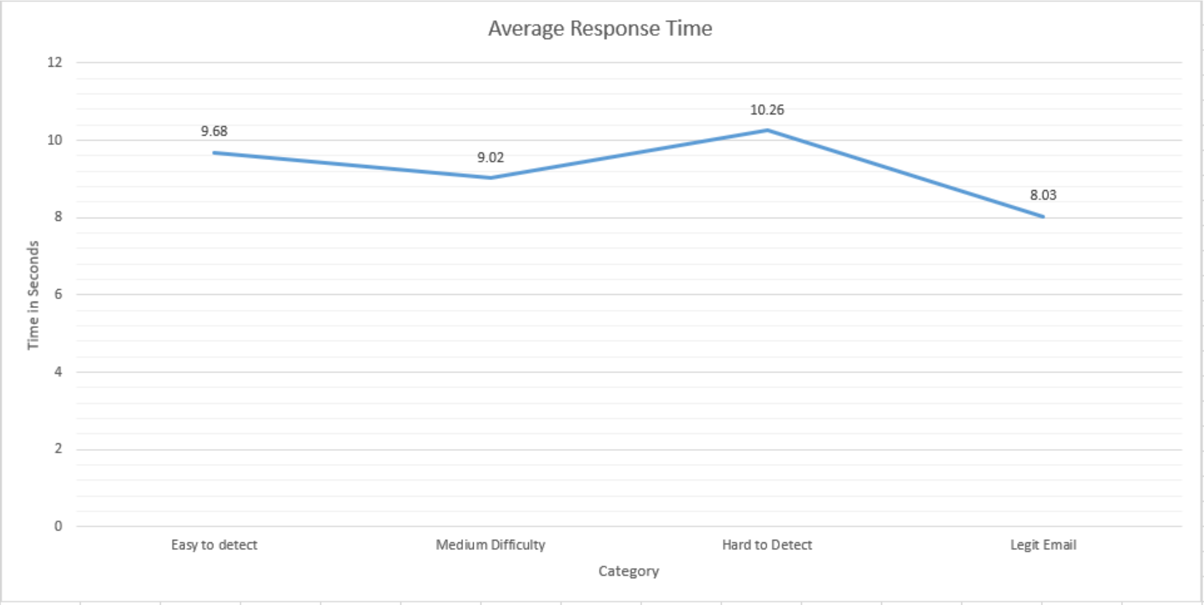


Figure 8: Average Response Time Across 4 Categories

As seen in figures 6 and 7, the response times are within the 13.3 seconds limit needed to read a 100 words email, with the Hard to detect category having the longest average response time of 10.26 seconds.

Compatibility

The artefact (browser extension) is compatible with all Chromium based web browsers, the accuracy and the response times were consistent between the tested browsers (Google Chrome, Brave, and Microsoft Edge).

Discussion

In the results section, we explored the accuracy and performance of the software, in this section, we will go into great detail of the results.

We evaluated the performance of the software both in accuracy and response times. The results indicate that the software successfully detected and identified the majority of the phishing attempts, with an overall score of 83.3%. The software was particularly effective in the Easy to detect, Hard to detect, and Legitimate emails categories with a 100% accuracy overall in these three categories, however, there is a huge discrepancy in the accuracy scores when it comes to the Medium Difficulty category where the software scored 33.3% accuracy, we believe that this is due to the mix of cues included in the design of the emails in that category.

The design of the emails in the Medium difficulty category intentionally included obvious cues such as urgency and authority, and a near perfect grammar and spelling, as well as legitimate URLs whether imbedded or in plaintext. Following the logic in the software's design, the voting system that was set in place to eliminate the false positives, in this case, returns a response that indicates a safe email, and the linguistic analysis of the near perfect grammar and spelling also returns a response that indicates a safe email. Further development of a refined voting system needs to be conducted to eliminate this issue.

On the other hand, the response time observed in the testing of the software, although might be considered long, is still within the timeframe that was calculated for the shortest average email (100 words in 13.3 seconds) by approximately 3.04 seconds.

These results indicate that the software preforms well in identifying a variety of phishing emails and within a timeframe that is within the safe limits.

Limitations

Although the overall accuracy score was relatively high with 83.3%, there were the instances discussed in the Medium difficulty category. As discussed earlier, we believe that the reason is because of the design of the emails in that category, which highlights a weak point in the software's design, that is, in the case of edge cases, one in particular might be advanced evasion techniques, such as using a URL shortener or encoding the URLs in a way that makes them appear legitimate and not get detected. Also, in case of a zero-day attack where neither VirusTotal nor GSB have identified the malicious code or URL, the voting system will rely solely on the linguistic analysis of the email, which will drive the possibility of false negatives and positives to a level where it becomes an annoyance rather than a solution.

Secondly, the response time, although is observed to be within the safe limits of reacting, is still high, a 25% margin is not enough time to use as contingency in case of any system delays or during high network traffic hours.

Future development will be needed to implement a solution to the edge cases, and another solution needs to be developed to improve the response time margins.

Implications and Future Work

The results have shown the effectiveness of the software in detecting and presenting an educational response to the end-user. As the development of the software progresses, it could potentially get to a point where it can be implemented in settings other than corporate environment, such as schools and personal email accounts.

Future work will explore the integration of machine learning algorithms to enhance the software's detection abilities and to eliminate any edge cases, also the possibility of developing an "in-house" solution to zero-day attacks instead of relying on third party software.

Another aspect of improvement that could be explored is the improvement of the linguistic analysis that is currently used in the software, an expansion of the intent detection vocabulary from one word description (i.e., "Negative"), or a simple analysis of what the request in the email is (i.e., “a request for information”) into a full paragraph analysis that might be able to detect edge cases or complex attacks that extends beyond the email itself and provide a better educational response to the end-user.

One more aspect that can be explored is the expansion of the software’s scope into detecting more than just phishing attempts, for instance, detecting and blocking ransomware attacks and overriding the end-user’s decisions to ignore the prompts provided to them when the analysis is giving a result that indicates a hundred percent detected attack attempt.

Conclusion

In conclusion, this thesis has investigated the cognitive reasons behind susceptibility to phishing attacks, and proposed a solution that combines the efforts of security awareness campaigns and the automated defence and detection systems that are already implemented by utilising a browser-based software that both detects and contextually educates end-users about the threats they are currently facing.

Studies discussed in this research have shown that susceptibility to phishing extends beyond the contents of an email or a message, they show that it is deeply connected to the cognitive behaviour of the end users, and the context that they are in at the time of receiving the email or message. The studies also showed the importance of security awareness campaigns and the effects of the contextual based awareness campaigns on the susceptibility to phishing attacks. It also briefly discussed the unforeseen negative side effects of overly automated solutions.

However, there is no denying that the proposed solution has its limitations, such as the potential for false positives and negatives, or the efficiency of the educational material that is delivered to the end users. Future work should focus on refining and fine-tuning the findings of the research to better design the software in a more effective and efficient way to allow for more accuracy and better generated educational material.

Moreover, the full extent of the proposed solution’s potential impact on the susceptibility to phishing can only truly be measured if the software is to be adopted and deployed by different organisations, to measure its impact on the level of awareness on employees.

References

- *Phishing activity trends reports* (2020-2022) *APWG*. Available at: https://apwg.org/trendsreports/ (Accessed: April 1, 2023).
- Carella, A., Kotsoev, M. and Truta, T.M. (2017) "Impact of security awareness training on phishing click-through rates," 2017 IEEE International Conference on Big Data (Big Data) [Preprint]. Available at: https://doi.org/10.1109/bigdata.2017.8258485.
- Dzulkifli, M.A. and Mustafar, M.F. (2013) The influence of colour on memory performance: A Review, The Malaysian journal of medical sciences : MJMS. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743993/#sec-9title (Accessed: April 2, 2023).
- National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743993/#sec-9title (Accessed: April 2, 2023).
- Kävrestad, J. et al. (2022) "Evaluation of contextual and game-based training for phishing detection," Future Internet, 14(4), p. 104. Available at: https://doi.org/10.3390/fi14040104.
- Lacey, D., Salmon, P. and Glancy, P. (2015) "Taking the bait: A systems analysis of phishing attacks," Procedia Manufacturing, 3, pp. 1109–1116. Available at: https://doi.org/10.1016/j.promfg.2015.07.185.
- McNamara, D.S. et al. (1996) "Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text," Cognition and Instruction, 14(1), pp. 1–43. Available at: https://doi.org/10.1207/s1532690xci1401_1.
- Mansfield-Devine, S. (2022) "IBM: Cost of a data breach," *Network Security*, 2022(8).

  Available at: https://doi.org/10.12968/s1353-4858(22)70049-9.
- Norman, D.A. (2021) The design of everyday things. New York, NY: Basic Books.
- Sawyer, B.D. and Hancock, P.A. (2018) "Hacking the human: The prevalence paradox in cybersecurity," Human Factors: The Journal of the Human Factors and Ergonomics Society, 60(5), pp. 597–609. Available at: https://doi.org/10.1177/0018720818780472.
- Steves, M., Greene, K. and Theofanos, M. (2020) "Categorizing human phishing difficulty: A phish scale," Journal of Cybersecurity, 6(1). Available at: https://doi.org/10.1093/cybsec/tyaa009.
- "Suspicion." Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/suspicion. Accessed 12 Apr. 2023.

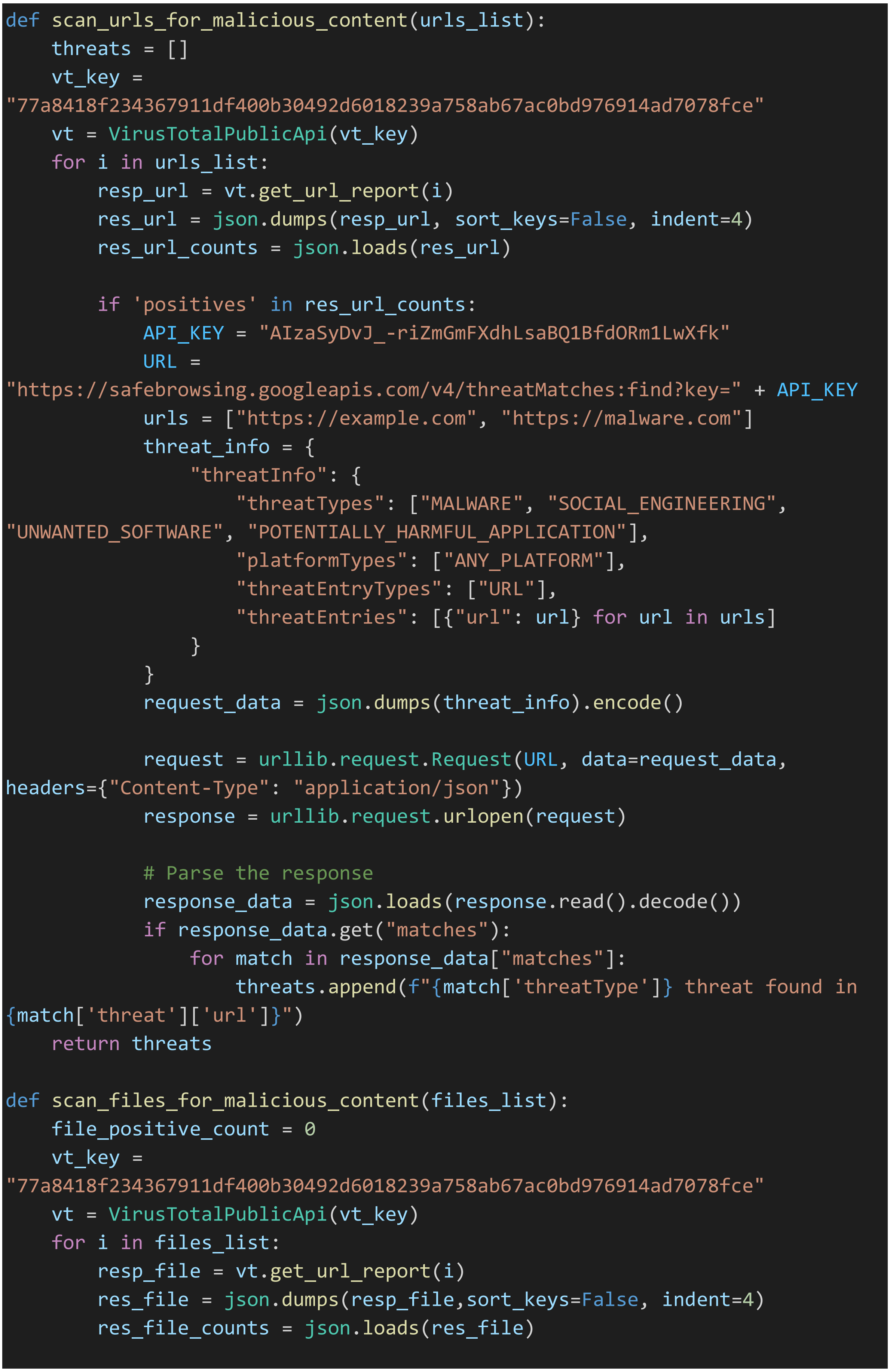
- Vishwanath, A., Harrison, B. and Ng, Y.J. (2016) “Suspicion, cognition, and automaticity model of phishing susceptibility,” Communication Research, 45(8), pp. 1146–1166. Available at: https://doi.org/10.1177/0093650215627483.
- Williams, E.J., Hinds, J. and Joinson, A.N. (2018) “Exploring susceptibility to phishing in the workplace,” International Journal of Human-Computer Studies, 120, pp. 1–13. Available at: https://doi.org/10.1016/j.ijhcs.2018.06.004.
- Carver, R.P. (1991) *Reading rate: A review of research and theory*. San Diego: Academic Press.

Appendices

Appendix A : Source Code

Python

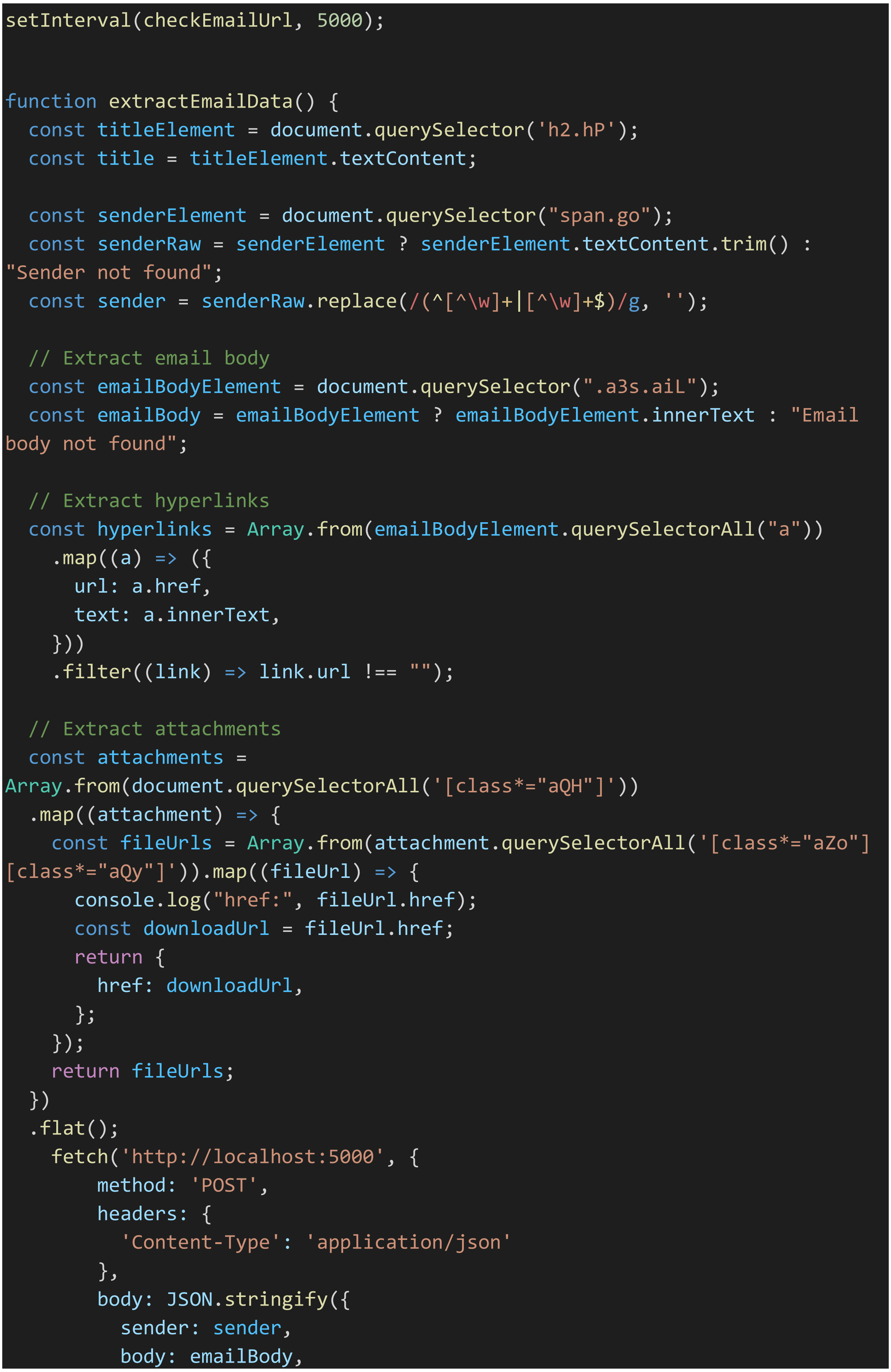
```
from flask import Flask, request as fRequest, jsonify
from flask_mysqldb import MySQL
from flask_cors import CORS, cross_origin
import openai
import json
from virus_total_apis import PublicApi as VirusTotalPublicApi
import urllib.request
import urllib.parse


app = Flask(__name__)
mysql = MySQL(app)
app.config['CORS_HEADERS'] = 'Content-Type'
CORS(app, resources={r"/api/*": {"origins": "*"}})

def extract_urls_from_email(anchor):
    urls_list = []
    for key in anchor:
        if 'url' in key:
            urls_list.append(key['url'])
    return urls_list

def extract_files_from_email(attachment):
    files_list = []
    for key in attachment:
        if 'href' in key:
            files_list.append(key['href'])
    return files_list
```



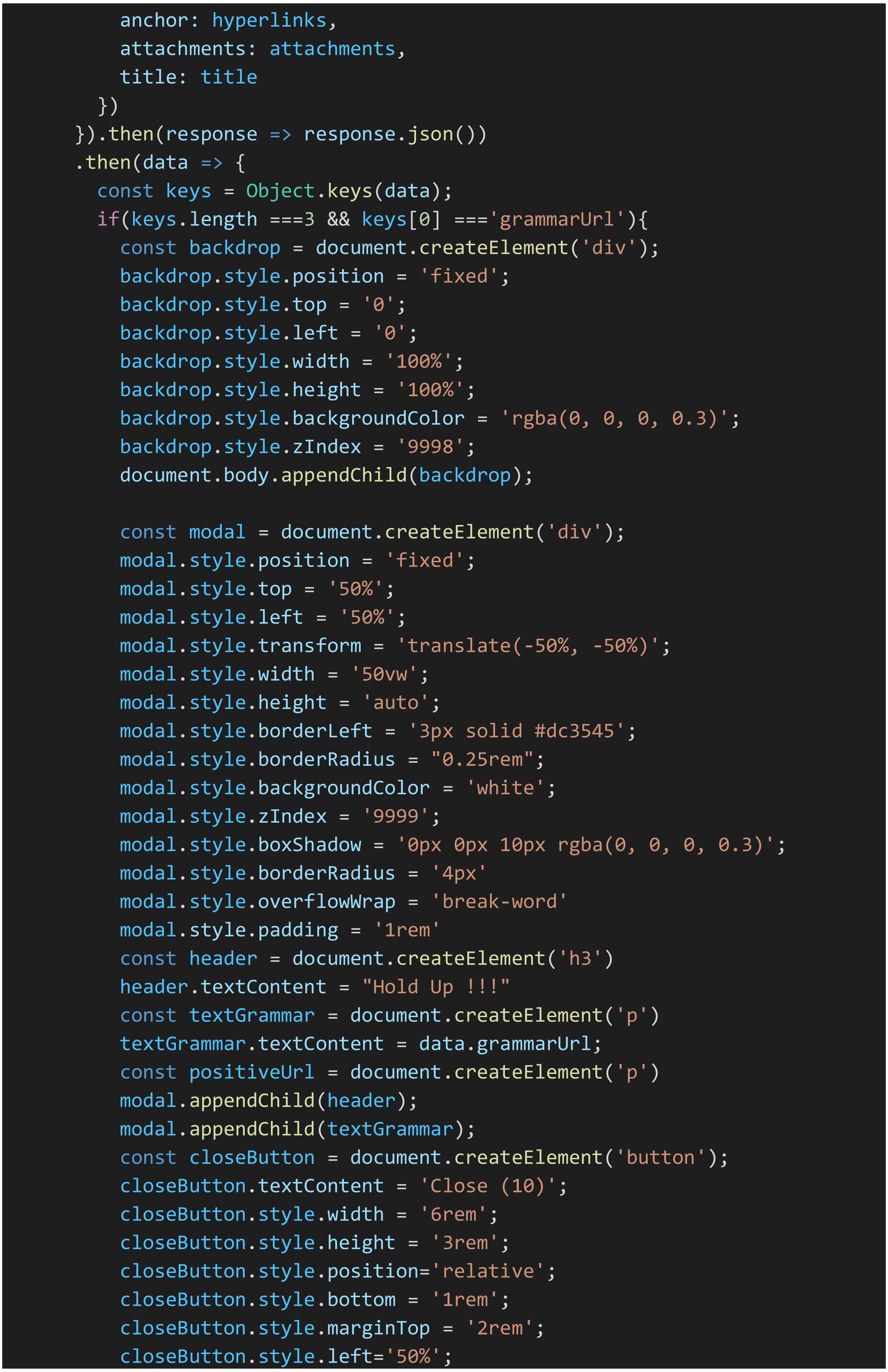
```
def scan_urls_for_malicious_content(urls_list):
    threats = []
    vt_key =
"77a8418f234367911df400b30492d6018239a758ab67ac0bd976914ad7078fce"
    vt = VirusTotalPublicApi(vt_key)
    for i in urls_list:
        resp_url = vt.get_url_report(i)
        res_url = json.dumps(resp_url, sort_keys=False, indent=4)
        res_url_counts = json.loads(res_url)

        if 'positives' in res_url_counts:
            API_KEY = "AIzaSyDvJ_-riZmGmFXdhLsaBQ1BfdORm1LwXfk"
            URL =
"https://safebrowsing.googleapis.com/v4/threatMatches:find?key=" + API_KEY
            urls = ["https://example.com", "https://malware.com"]
            threat_info = {
                "threatInfo": {
                    "threatTypes": ["MALWARE", "SOCIAL_ENGINEERING",
"UNWANTED_SOFTWARE", "POTENTIALLY_HARMFUL_APPLICATION"],
                    "platformTypes": ["ANY_PLATFORM"],
                    "threatEntryTypes": ["URL"],
                    "threatEntries": [{"url": url} for url in urls]
                }
            }
            request_data = json.dumps(threat_info).encode()

            request = urllib.request.Request(URL, data=request_data,
headers={"Content-Type": "application/json"})
            response = urllib.request.urlopen(request)

            # Parse the response
            response_data = json.loads(response.read().decode())
            if response_data.get("matches"):
                for match in response_data["matches"]:
                    threats.append(f"{match['threatType']} threat found in
{match['threat']['url']}")
    return threats

def scan_files_for_malicious_content(files_list):
    file_positive_count = 0
    vt_key =
"77a8418f234367911df400b30492d6018239a758ab67ac0bd976914ad7078fce"
    vt = VirusTotalPublicApi(vt_key)
    for i in files_list:
        resp_file = vt.get_url_report(i)
        res_file = json.dumps(resp_file,sort_keys=False, indent=4)
        res_file_counts = json.loads(res_file)
        if 'positives' in res_file_counts:
            if res_file_counts['positives'] > file_positive_count:
                file_positive_count = res_file_counts['positives']
    return file_positive_count

def generate_text_prompt(sender, title, mainBody):
    prompt = f"""
    Please analyze the following email and provide your opinion on whether
it appears to be a phishing email or not. The email contains the following
information:
    Sender: {sender}
    Title: {title}
    Body: {mainBody}
    Based on the contents of the email, please provide your opinion on
whether it appears to be a phishing attempt or not. If you believe the email
may be a phishing attempt. Phishing emails often contain malicious links or
files that can harm your computer or steal your personal information. If
you're unsure about the legitimacy of an email, you can always reach out to
your organization's IT department for assistance.
    Please provide an explanation of why you believe the email is or is not
a phishing email and tell me why you.
    """
    return prompt

@app.route('/', methods=['POST'])
@cross_origin(origin='localhost', headers=['Content-Type', 'Authorization'])
def analyse():
    vt_key =
"77a8418f234367911df400b30492d6018239a758ab67ac0bd976914ad7078fce"
    vt = VirusTotalPublicApi(vt_key)
    OA_key = "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"
    openai.api_key = OA_key
    model = "text-davinci-003"

    body = fRequest.get_json('body')
    title = body['title']
    sender = body['sender']
    mainBody = body['body']
    anchor = body['anchor']
    attachment = body['attachments']

    urls_list = extract_urls_from_email(anchor)
    threats = scan_urls_for_malicious_content(urls_list)

    files_list = extract_files_from_email(attachment)
    file_positive_count = scan_files_for_malicious_content(files_list)

    prompt = generate_text_prompt(sender, title, mainBody)
    generate_text = openai.Completion.create(engine=model, prompt=prompt,
max_tokens=512, n=1, stop=None, temperature=0.7, echo=False, best_of=3)
    grammar_check = generate_text.choices[0].text.strip()

    sentiment_prompt = f'Please analyze this text {grammar_check} and
classify its sentiment as either Positive or Negative in one word.'
    sentiment_analysis =
openai.Completion.create(engine=model,prompt=sentiment_prompt,temperature=0.
5,max_tokens = 512,n=1,stop=None,frequency_penalty=0,presence_penalty=0)
    sentiment = sentiment_analysis.choices[0].text.strip()


    if len(threats) > 0 or file_positive_count > 0 or sentiment
=='Negative':
        return jsonify({'grammarUrl': grammar_check, 'textUrl': urls_list,
'textFile': files_list})
    elif len(threats) == 0 and file_positive_count == 0 and sentiment
=='Negative':
        return jsonify({'grammarUrl': grammar_check, 'textUrl': urls_list,
'textFile': files_list})
    elif (len(threats) > 0 or file_positive_count > 0) and sentiment
=='Positive':
        return jsonify({'grammarUrl': grammar_check, 'textUrl': urls_list,
'textFile': files_list})
    else:
        return jsonify({'grammarUrl': grammar_check})


if __name__ == '__main__':
    app.run(debug=True)
```

JavaScript

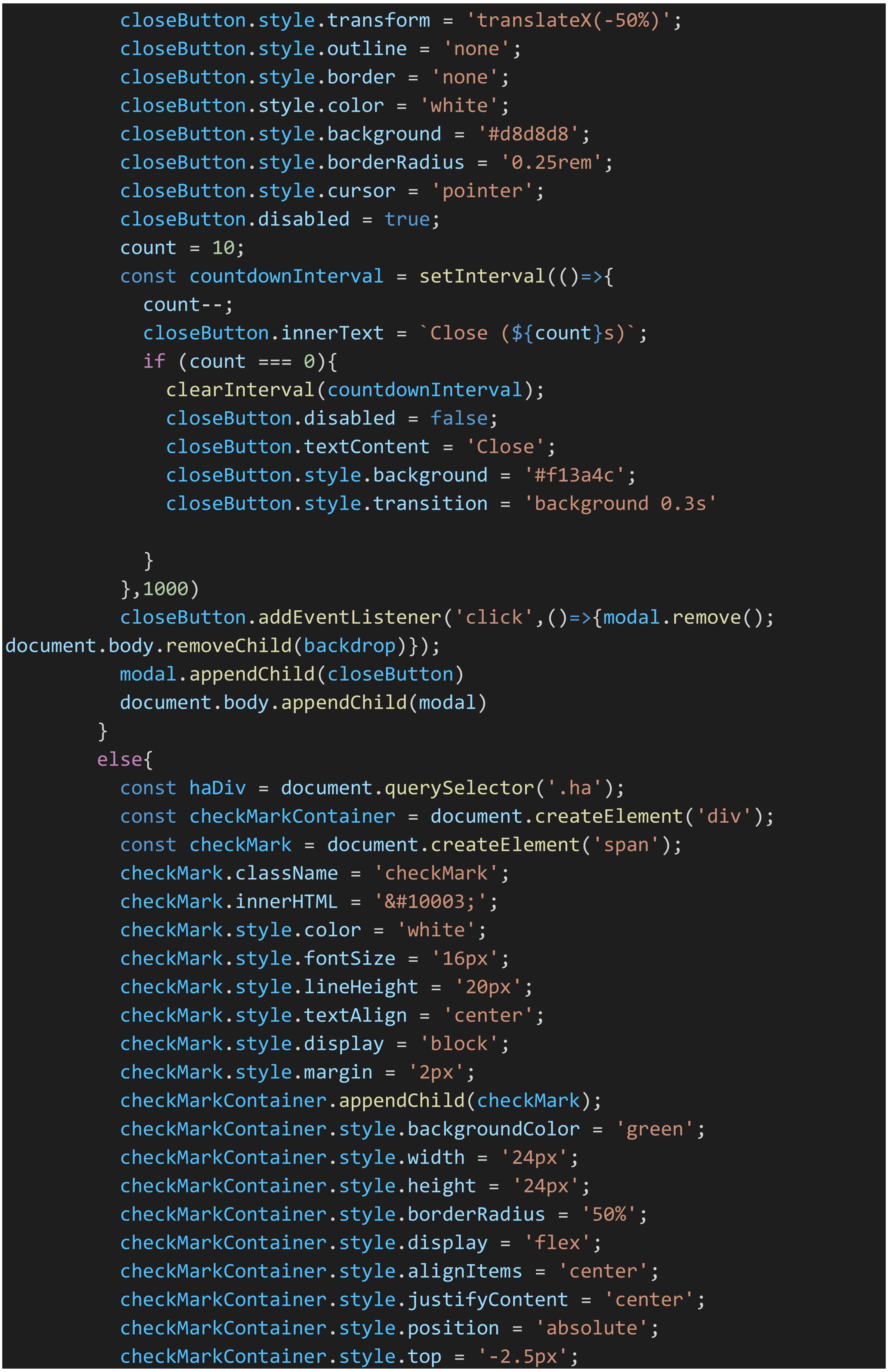
```
let initialUrl = location.href;

function checkEmailUrl() {
  if (location.href !== initialUrl) {
    console.log("URL has changed. Calling extractEmailData()...");
    extractEmailData();
    initialUrl = location.href;
  }
}

console.log("Initial check...");
checkEmailUrl();
```



```
setInterval(checkEmailUrl, 5000);


function extractEmailData() {
  const titleElement = document.querySelector('h2.hP');
  const title = titleElement.textContent;

  const senderElement = document.querySelector("span.go");
  const senderRaw = senderElement ? senderElement.textContent.trim() :
"Sender not found";
  const sender = senderRaw.replace(/(^[^\w]+|[^\w]+$)/g, '');

  // Extract email body
  const emailBodyElement = document.querySelector(".a3s.aiL");
  const emailBody = emailBodyElement ? emailBodyElement.innerText : "Email
body not found";

  // Extract hyperlinks
  const hyperlinks = Array.from(emailBodyElement.querySelectorAll("a"))
    .map((a) => ({
      url: a.href,
      text: a.innerText,
    }))
    .filter((link) => link.url !== "");

  // Extract attachments
  const attachments =
Array.from(document.querySelectorAll('[class*="aQH"]'))
  .map((attachment) => {
    const fileUrls = Array.from(attachment.querySelectorAll('[class*="aZo"]
[class*="aQy"]')).map((fileUrl) => {
      console.log("href:", fileUrl.href);
      const downloadUrl = fileUrl.href;
      return {
        href: downloadUrl,
      };
    });
    return fileUrls;
  })
  .flat();
    fetch('http://localhost:5000', {
        method: 'POST',
        headers: {
          'Content-Type': 'application/json'
        },
        body: JSON.stringify({
          sender: sender,
          body: emailBody,
```



```
        anchor: hyperlinks,
        attachments: attachments,
        title: title
      })
    }).then(response => response.json())
    .then(data => {
      const keys = Object.keys(data);
      if(keys.length ===3 && keys[0] ==='grammarUrl'){
        const backdrop = document.createElement('div');
        backdrop.style.position = 'fixed';
        backdrop.style.top = '0';
        backdrop.style.left = '0';
        backdrop.style.width = '100%';
        backdrop.style.height = '100%';
        backdrop.style.backgroundColor = 'rgba(0, 0, 0, 0.3)';
        backdrop.style.zIndex = '9998';
        document.body.appendChild(backdrop);

        const modal = document.createElement('div');
        modal.style.position = 'fixed';
        modal.style.top = '50%';
        modal.style.left = '50%';
        modal.style.transform = 'translate(-50%, -50%)';
        modal.style.width = '50vw';
        modal.style.height = 'auto';
        modal.style.borderLeft = '3px solid #dc3545';
        modal.style.borderRadius = "0.25rem";
        modal.style.backgroundColor = 'white';
        modal.style.zIndex = '9999';
        modal.style.boxShadow = '0px 0px 10px rgba(0, 0, 0, 0.3)';
        modal.style.borderRadius = '4px'
        modal.style.overflowWrap = 'break-word'
        modal.style.padding = '1rem'
        const header = document.createElement('h3')
        header.textContent = "Hold Up !!!"
        const textGrammar = document.createElement('p')
        textGrammar.textContent = data.grammarUrl;
        const positiveUrl = document.createElement('p')
        modal.appendChild(header);
        modal.appendChild(textGrammar);
        const closeButton = document.createElement('button');
        closeButton.textContent = 'Close (10)';
        closeButton.style.width = '6rem';
        closeButton.style.height = '3rem';
        closeButton.style.position='relative';
        closeButton.style.bottom = '1rem';
        closeButton.style.marginTop = '2rem';
        closeButton.style.left='50%';
```



```
          closeButton.style.transform = 'translateX(-50%)';
          closeButton.style.outline = 'none';
          closeButton.style.border = 'none';
          closeButton.style.color = 'white';
          closeButton.style.background = '#d8d8d8';
          closeButton.style.borderRadius = '0.25rem';
          closeButton.style.cursor = 'pointer';
          closeButton.disabled = true;
          count = 10;
          const countdownInterval = setInterval(()=>{
            count--;
            closeButton.innerText = `Close (${count}s)`;
            if (count === 0){
              clearInterval(countdownInterval);
              closeButton.disabled = false;
              closeButton.textContent = 'Close';
              closeButton.style.background = '#f13a4c';
              closeButton.style.transition = 'background 0.3s'

            }
          },1000)
          closeButton.addEventListener('click',()=>{modal.remove();
document.body.removeChild(backdrop)});
          modal.appendChild(closeButton)
          document.body.appendChild(modal)
        }
        else{
          const haDiv = document.querySelector('.ha');
          const checkMarkContainer = document.createElement('div');
          const checkMark = document.createElement('span');
          checkMark.className = 'checkMark';
          checkMark.innerHTML = '&#10003;';
          checkMark.style.color = 'white';
          checkMark.style.fontSize = '16px';
          checkMark.style.lineHeight = '20px';
          checkMark.style.textAlign = 'center';
          checkMark.style.display = 'block';
          checkMark.style.margin = '2px';
          checkMarkContainer.appendChild(checkMark);
          checkMarkContainer.style.backgroundColor = 'green';
          checkMarkContainer.style.width = '24px';
          checkMarkContainer.style.height = '24px';
          checkMarkContainer.style.borderRadius = '50%';
          checkMarkContainer.style.display = 'flex';
          checkMarkContainer.style.alignItems = 'center';
          checkMarkContainer.style.justifyContent = 'center';
          checkMarkContainer.style.position = 'absolute';
          checkMarkContainer.style.top = '-2.5px';
          checkMarkContainer.style.right = '-30px';
          checkMarkContainer.style.margin = '5px';
          haDiv.style.position = 'relative';
          haDiv.appendChild(checkMarkContainer);
          var tooltip = document.createElement('div');
          tooltip.style.position = 'absolute';
          tooltip.style.zIndex = '999';
          tooltip.style.backgroundColor = 'rgb(40,167,69)';
          tooltip.style.color = '#0d3717';
          tooltip.style.borderRadius = '6px';
          tooltip.style.padding = '10px';
          tooltip.style.opacity = '0';
          tooltip.style.border = "1px solid #28a745";
          tooltip.style.transition = 'opacity 0.3s';
          tooltip.style.maxWidth = '600px';
          tooltip.style.wordBreak = 'break-word';
          document.body.appendChild(tooltip);
          checkMark.addEventListener('mouseover',function(event){
          tooltip.textContent = data.grammarUrl;
          tooltip.style.top = ((event.clientY+10)+'px');
          tooltip.style.left = ((event.clientX+10) + 'px');
          tooltip.style.opacity = '1';
        });
          checkMark.addEventListener('mouseout',function(){
            tooltip.style.opacity = '0';
          })
        }
      })
}

setTimeout(extractEmailData,2500)
```

Manifist.JSON

```
{
    "manifest_version": 3,
    "name": "Secure Inbox",
    "version": "1.0",
    "description": "An extension to analyise the contents of the email, to protect against malicious activity",
      "author" : "Amjed Ashour",
      "permissions" : [
        "activeTab"
    ],
    "icons":{
      "16":"static/16.png",
      "32":"static/32.png",
      "48":"static/48.png",
      "128":"static/128.png"
    },

  "web_accessible_resources": [{
  "resources": ["jquery.min.js",
        "templates/modal.html",
        "dialog-polyfill.js"],
  "matches": ["<all_urls>"]
}],
    "content_scripts":[
        {
            "matches": [
                "*://mail.google.com/*",
                "*://outlook.live.com/*",
                "*://mail.yahoo.com/*",
                "*://www.icloud.com/*",
                "*://www.apple.com/mail/*",
                "*://mail.aol.com/*"
              ],
              "js":["scripts.js"]
        }
    ]

}
```

Appendix B: Test Emails

Easy to Detect
Title: Urgent- Your Account Has Been Compromised **Body**: Dear Customer, We have detected a suspicious activity on your account. To secure your account click on the link to provide your login details: https://www.aiforums.com/credintials-update
Title: Academic misconduct **Body**:

Dear Student, Some students have been exploiting the advancement of AI by submitting an AI generated paper for their finals. Our systems detected that you have submitted an AI generated paper, therefore you have been suspended until we conclude our investigation, please login to your account to check the date of your disciplinary hearing: https://www.aiforums.com/disciplinary-hearing Best wishes, Teaching staff.
Title: Update Your Profile and Win Big!! **Body**: We are happy tp announce our “up-to-date” lottery. Update your accounts information here: https://www.aiforums.com/account-update to have a chance at winning 10000 dollars. Goodluck, Social media team.
Medium Difficulty
Title: Urgent Alert: Your Account has been compromised **Body**: Dear Amjed, We regret to inform you that we have detected some suspicious activity on your account. Our security team has identified that your account has been compromised and is at risk of being misused. We take the security of our customers’ information seriously, and we want to ensure that your account remains safe and secure. To protect your account, we urge you to take immediate action by clicking the link below to update your login credentials. Please do not ignore this message, as failure to take action may result in the unauthorized use of your account. https://www.aiforums.com/credintials-update We advise you to exercise caution when providing personal information and to use strong and unique passwords to prevent any future unauthorized access. Thank you for your cooperation in this matter. Sincerely, The Management Team.
Title: Academic Misconduct **Body**: It has come to our attention that a few students have been resorting to unethical practices by submitting papers generated by artificial intelligence for their final assessments. We regret to inform you that our systems have detected the use of an AI-generated paper in your submission.

As a consequence, we have temporarily suspended your academic activities pending the completion of our investigation into this matter. We take academic integrity very seriously and would like to ensure that our students adhere to the highest standards of ethical conduct.

We request that you log in to your account here to check the date of your disciplinary hearing, where you will have the opportunity to present your case. We assure you that we will conduct a thorough investigation and take appropriate actions based on our findings.

Thank you for your understanding and cooperation.

Sincerely,

Teaching staff.

Title: Participate in Our Profile Update Lottery and Win Big!

Body: Dear Esteemed User,

We are delighted to inform you of our latest promotion, the "Up-to-date" lottery, where you stand a chance to win a grand prize of $10,000.

To participate in this exciting opportunity, we encourage you to update your account information by clicking on the link below:

https://www.aiforums.com/account-update

By updating your profile, you not only get the chance to win a significant cash prize, but you also ensure that your account remains current and secure. It's a win-win situation!

Please note that the promotion is open for a limited time only. Therefore, we urge you to update your account information as soon as possible and stand a chance to win big!

Thank you for being a valued member of our community. We wish you the best of luck in the lottery.

Sincerely,

The Social Media Team.

Hard to Detect

Title: Account Security Notice: Action Required

Body: Dear Amjed,

We recently identified some unusual activity on your account. As a precautionary measure, our security team has temporarily limited certain actions on your account to prevent any potential misuse. We take the security of our customers' information very seriously and want to ensure that your account remains safe and secure.

To restore your account to its full functionality, we kindly ask you to follow the link below to verify your login credentials. This is an essential step to confirm your identity and safeguard your account.

https://www.legitimate-website.com/account-security-verification

Please note that this link will expire in 48 hours. If you need any assistance or have questions regarding this matter, do not hesitate to contact our support team at support@legitimate-website.com.

We also encourage you to use strong and unique passwords for your online accounts and to remain vigilant when sharing personal information online.

Thank you for your cooperation and understanding.

Sincerely,

The Management Team.

Title: Academic Integrity Concern: Please Review Your Submission

Body: Dear Student,

We have noticed some irregularities in the content of the paper submitted for your recent final assessment. Our systems have detected similarities with AI-generated content, which raises concerns about academic integrity. At our institution, we maintain the highest standards of ethical conduct, and it is crucial that all students adhere to these principles.

In order to clarify this matter, we kindly ask you to review your submission and provide any additional information or context that can help us better understand the situation. Please log in to your account at the following link to access your submission and respond to our inquiry:

https://www.legitimate-education-website.com/review-submission

We assure you that we will carefully consider any information you provide and will conduct a fair and unbiased evaluation of the case. We appreciate your cooperation and understanding during this process.

Thank you, and we look forward to your prompt response.

Sincerely,

Teaching Staff.

Title: Celebrate Our 10th Anniversary with a Chance to Win a $10,000 Prize

Body: Dear Esteemed User,

As we celebrate the 10th anniversary of our platform, we want to show our appreciation to our valued users like you. To mark this special occasion, we are excited to announce the "10th Anniversary Giveaway," where you stand a chance to win a grand prize of $10,000.

To participate in this exciting opportunity, simply review and update your account information by visiting the link below:

https://www.legitimate-social-media.com/account-update

By updating your profile, not only do you get the chance to win a significant cash prize, but you also ensure that your account remains current and secure. It's a win-win situation!

Please note that the giveaway is open for a limited time only. Therefore, we encourage you to update your account information as soon as possible to enter the contest and join us in celebrating our 10th anniversary.

Thank you for being a valued member of our community. We wish you the best of luck in the giveaway and look forward to many more years together! Sincerely, The Social Media Team.
Legitimate Emails
Title: Security Update: Recent Data Breach and Account Verification **Body**: Dear Amjed, We regret to inform you that our platform has recently experienced a data breach. While our security team has contained the situation, we are taking precautionary measures to ensure the safety of our users' accounts. As part of these measures, we kindly ask you to verify your account information to confirm that your account remains secure. Please login to your account and change your password and contact information. We understand the importance of safeguarding your personal information and are committed to taking all necessary steps to protect it. We apologize for any inconvenience this may cause and appreciate your cooperation in maintaining the security of your account. If you have any questions or concerns, please feel free to reach out to our support team at support@aiforums.com. Thank you for choosing aiforums Sincerely, Security Team at aiforums.
Title: Academic Integrity Concern and Disciplinary Hearing Notification **Body**: Dear Student, We have received your recent coursework submission and during our evaluation process, we have identified some concerns regarding academic integrity. It appears that there may have been a violation of our institution's academic honesty policies in your submitted work regarding using Artificial Intelligence generated papers. As a result, we have scheduled a disciplinary hearing to address these concerns and provide you with an opportunity to present your case. The details of your disciplinary hearing, including the date and time, can be found by logging into your account and checking the Academic Integrity page. Our institution upholds the highest standards of academic honesty, and we expect all students to adhere to these principles. We take any potential breach of academic integrity seriously and will conduct a thorough investigation into the matter. If you have any questions or need further information about the hearing or our academic integrity policies, please do not hesitate to contact your course instructor or visit your course dashboard. Sincerely, Academic Integrity Board.

Title: 10th Anniversary Celebration: Enter Our $10,000 Lottery!

Body: Dear Amjed,

We are excited to celebrate the 10th anniversary of our platform, and we want to share this milestone with our valued users like you. To mark this special occasion, we are delighted to announce a $10,000 lottery exclusively for our community members.

To participate in the "10th Anniversary Giveaway," all you need to do is review and update your account information Under the Profile Settings section.

Updating your profile not only enters you into the giveaway for a chance to win the grand prize of $10,000, but it also helps us ensure that your account remains current and secure.

Please note that the giveaway is open for a limited time only, so be sure to update your account information as soon as possible to secure your entry.

Thank you for being a valued member of our community. We wish you the best of luck in the giveaway and look forward to celebrating many more milestones together!

Sincerely,

The Social Media Team.