POSTER: Email Summarization to Assist Users in Phishing Identification

Amir Kashapov
akas0005@student.monash.edu
Monash University, Australia

Tingmin Wu
tina.wu1@monash.edu
Monash University, CSIRO's Data61, Australia

Sharif Abuadbba
sharif.abuadbba@data61.csiro.au
CSIRO's Data61, Australia

Carsten Rudolph
carsten.rudolph@monash.edu
Monash University, Australia

Abstract

Cyber-phishing attacks recently became more precise, targeted, and tailored by training data to activate only in the presence of specific information or cues. They are adaptable to a much greater extent than traditional phishing detection. Hence, automated detection systems cannot always be 100% accurate, increasing the uncertainty around expected behavior when faced with a potential phishing email. On the other hand, human-centric defence approaches focus extensively on user training but face the difficulty of keeping users up to date with continuously emerging patterns. Therefore, advances in analyzing the content of an email in novel ways along with summarizing the most pertinent content to the recipients of emails is a prospective gateway to furthering how to combat these threats. Addressing this gap, this work leverages transformer-based machine learning to (i) analyze prospective psychological triggers, to (ii) detect possible malicious intent, and to (iii) create representative summaries of emails. We then amalgamate this information and present it to the user to allow them to (i) easily decide whether the email is "phishy" and (ii) self-learn advanced malicious patterns.

CCS Concepts

• **Security and privacy → Network security**; • **Computing methodologies → Natural language processing**.

Keywords

Phishing, Email, Machine Learning, Summarization

ACM Reference Format:
Amir Kashapov, Tingmin Wu, Sharif Abuadbba, and Carsten Rudolph. 2022. POSTER: Email Summarization to Assist Users in Phishing Identification. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30-June 3, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3488932.3527292

1 Introduction

Phishing is a type of cyber-attack whereby criminals design seemingly authentic emails with the intent of tricking users into giving up private, confidential, and/or sensitive information such as login passwords and financial data. Artificial Intelligence-based automated phishing detection methods are limited as Machine Learning (ML) models can only identify the malicious patterns they have been trained on and detection by heuristics-based techniques produce high false positive rates [3]. In today's highly connected world, users spend considerable time sending and reading both work and personal emails, and it is thus hard for users to pay close attention to every email. Consequently, they can fall victim to advanced deceptive phishing techniques that bypass security filters.

Recent work shifts the focus from automated phishing detection to detection support to assist users in making their own judgements [2], since automated approaches are not always 100% accurate and real-time support such as warnings can effectively change risky behaviors. Presenting users with security indicators' information enables human strength in capturing abnormal behaviors, such as contextual awareness.A human-centric solution using autonomous ML agents to aid judgment can therefore be a crucial step in the right direction. Despite security warnings and modern means of educating users about phishing, current detection support and training still cannot deal with sophisticated malicious emails in real time. Therefore, condensing important email information relevant to phishing and rationalising human efforts are important objectives to deal with new types of attacks.

Most of the existing works on phishing detection support heavily rely on the legitimacy of URLs [1, 2]. However, explanations of URL features are easy to understand by general users, e.g., website rank, hostname, and domain popularity. More importantly, phishing attacks have evolved to escape URL detection by leveraging social engineering. In particular, criminals can generate links through third-party services or a plaintext with a response request. Motivated by that, we aim to address the challenges of improving the readability and effectiveness of generated information on phishing emails that URL-based detection methods cannot detect.

Therefore, the objective of this work is to design and develop a human-centric notification system based on both emails and their contexts using Natural Language Processing (NLP) methods to help users accurately identify phishing attempts. The intention is to outline useful information based only on the email and the contextual knowledge that machines can utilize - to be precise, this includes the summary of an email, the emotions associated with an email and the intents of the sender that may be relevant to evoking self-sabotaging actions.

ASIA CCS '22, May 30-June 3, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9140-5/22/05.
https://doi.org/10.1145/3488932.3527292

1.1 Contributions

We present a novel, human-centric mechanism to combat one of the most prevalent cybercrimes of the world today - email phishing. Our contributions to that end can be summarized as follows:

(1) We propose a system to pipeline emails through ML models and to generate a human-centric summary report on useful criteria for the user to identify phishing.
(2) We leverage and investigate the utility of the summarization capability of the latest transformer-based ML models on emails for phishing analysis.
(3) We leverage and investigate the utility of the emotion classification capability of transformer-based ML models on emails in the context of "cognitive triaging" to detect potential psychological triggers.
(4) We leverage and investigate the utility of the intent analysis capability of transformer-based ML models on emails to determine whether malicious intent can be accurately "cherry-picked".

2 System Design

We identified three critical components in particular as utilities of state-of-the-art ML models that could indicate potential phishing emails, and, as a result, we build our system to combine them. The system design thus incorporates three fundamental pipelines: extractive summarization, emotion classification, and intent analysis. We next explain those three pipelines.

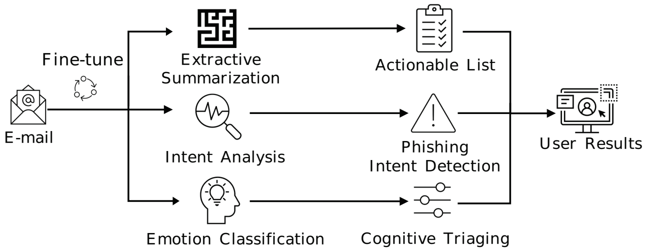


Figure 1: System Design Overview

Extractive Summarization. Phishing emails can be overwhelmingly long to deceive the recipients to overlook the phishing components and to focus on urgency and/or emotions. To emphasize the most important context, we use the summarization capability of modern ML models. Informing the user of the most important contents of the email and of the potentially desired actionable items evoked by the sender could allow them to examine the primary purpose of the email without any superfluous or distracting information clouding their judgement. For instance, if a small part of the email evokes a potential victim to click on a harmful link whereas the others are merely warnings of what would occur if they don't, an ideal summarization branch/pipeline would immediately point this out as as irregular as part of the user results.

Emotion Classification. The multi-class emotion classification branch is inspired by the brilliant paper on cognitive triaging [6]. There are six possible "cognitive triggers" we detect in phishing emails - *Reciprocity, Consistency, Social Proof, Authority, Liking, and Scarcity* [6]. The definition of these terms are, respectively, as follows: *Reciprocity* - Tendency to feel obliged to repay favors from others. *Consistency* - Tendency to behave in a way consistent with past decisions and behaviors. *Social Proof* - Tendency to reference the behavior of others to guide one's own actions. *Authority* - Tendency to obey people in authoritative positions. *Liking* - Preference for saying "yes" to the requests of people one knows and likes. *Scarcity* - Tendency to assign more value to items and opportunities when their availability is limited [6].

Intent Analysis. This analysis filters and detects potentially malicious intent in the form of actionable and identifiable items that can be detected through the use of ML models. The aim of this branch is that a recipient gets a clearer picture as to whether an email is harmful when highlighted and contextualized suspicious portions of an email are taken into consideration with information from the other branches. Highlighting potential phishing intent could also serve to train recipients to develop a "feel" for what phishing intents and signals look like.

3 Preliminary results

3.1 Data collection and pre-processing

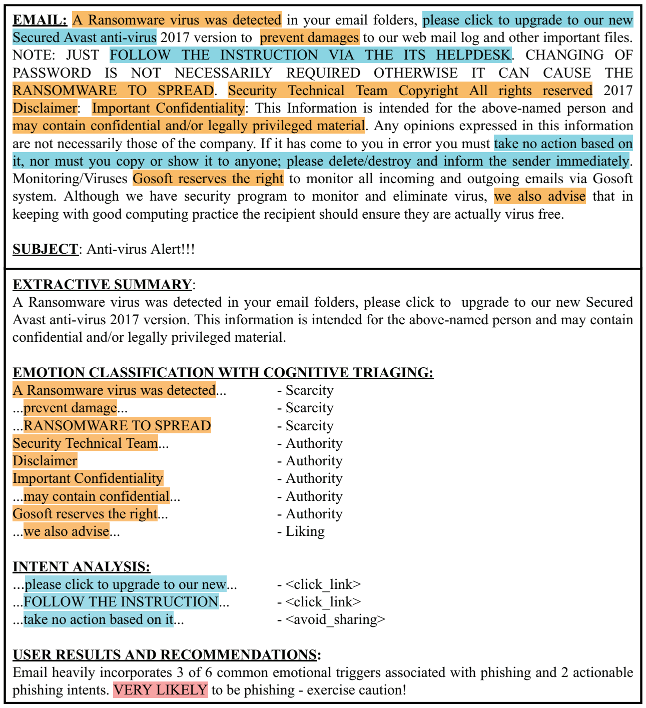
The first experimental setup step was to compile email datasets which include the Cambridge phishing dataset, the Cornell phish bowl, the Enron email corpus of benign emails, the millersmiles.co.uk phishing dataset, and the Nazario phishing dataset. The above were chosen as they were collected by reputable researchers to create an accurate representation of phishing contexts or of benign correspondence in emails - in fact, all datasets in question compile real-life emails. The corpora contain 41446, 1757, 252721, 33080, and 946 emails respectively, and were all utilized to create random experiment samples of varying sizes in a variety of contexts - early evaluation sample sizes numbered 500 or more emails followed by final evaluation samples numbering 5000 or more for all branches.

For all three branches, pre-processing separated the email body from the rest of the email and removed HTML tags if they were present. In addition, pre-processing for the intent analysis and emotion classification pipelines also included the splitting of email bodies into separate sets of words each with an associated set of labels such as relevant emotions or intent.

3.2 Implementation

Recently, the advent of modern leading ML models has been propagated and bolstered largely thanks to the release of the "Transformer" by researchers at Google [7]. As a consequence of its release, current leading models in terms of the universality of their applicability in the ML space such as T5 [5], XLNet [8], and BERT [4] all integrate the Transformer to varying degrees. Such universality means that they can be applied in a variety of different contexts, including emails, to obtain impressive results.

Extractive Summarization. We build a summarization algorithm that leverages recent transformer-based models: T5 [5], XLNet [8], and BERT [4]. T5 has produced the most promising results and, as a result, it has been adopted for this task. T5 has the advantage that it does not need to be fine-tuned for individual tasks by design. As a result, the implementation of the task and purpose with this model has been somewhat trivial. Whereas the utility of achieving the above may be evident, the precise manner in which content is to be summarized in an ideal scenario is of course open to interpretation since there are no objective criteria. As a result, fractional limits have been set with regards to the word length of a generated summary for any given email for examination in addition to hard-set limits such as 25, 50, or 100 words. Closer examination and some metric such as cross-validation is potentially needed to decide on how to best ascertain what word limit the summary should have - through admittedly looser criteria such as subjectively judging grammar, readability, and the ratio of summary length to essential context, the most promising result has been to set the maximal length of a summary to be a fraction of one fifth of the length of the original email.



EMAIL: A Ransomware virus was detected in your email folders, please click to upgrade to our new Secured Avast anti-virus 2017 version to prevent damages to our web mail log and other important files. NOTE: JUST FOLLOW THE INSTRUCTION VIA THE ITS HELPDESK. CHANGING OF PASSWORD IS NOT NECESSARILY REQUIRED OTHERWISE IT CAN CAUSE THE RANSOMWARE TO SPREAD. Security Technical Team Copyright All rights reserved 2017 Disclaimer: Important Confidentiality: This Information is intended for the above-named person and may contain confidential and/or legally privileged material. Any opinions expressed in this information are not necessarily those of the company. If it has come to you in error you must take no action based on it, nor must you copy or show it to anyone; please delete/destroy and inform the sender immediately. Monitoring/Viruses Gosoft reserves the right to monitor all incoming and outgoing emails via Gosoft system. Although we have security program to monitor and eliminate virus, we also advise that in keeping with good computing practice the recipient should ensure they are actually virus free.

SUBJECT: Anti-virus Alert!!!

EXTRACTIVE SUMMARY:
A Ransomware virus was detected in your email folders, please click to upgrade to our new Secured Avast anti-virus 2017 version. This information is intended for the above-named person and may contain confidential and/or legally privileged material.

EMOTION CLASSIFICATION WITH COGNITIVE TRIAGING:

A Ransomware virus was detected...	- Scarcity
...prevent damage...	- Scarcity
...RANSOMWARE TO SPREAD	- Scarcity
Security Technical Team...	- Authority
Disclaimer	- Authority
Important Confidentiality	- Authority
...may contain confidential...	- Authority
Gosoft reserves the right...	- Authority
...we also advise...	- Liking

INTENT ANALYSIS:

...please click to upgrade to our new...	- <click_link>
...FOLLOW THE INSTRUCTION...	- <click_link>
...take no action based on it...	- <avoid_sharing>

USER RESULTS AND RECOMMENDATIONS:
Email heavily incorporates 3 of 6 common emotional triggers associated with phishing and 2 actionable phishing intents. VERY LIKELY to be phishing - exercise caution!

Figure 2: An example of a phishing email, results from the three pipelines, and the final system-generated user results.

Emotion Classification. We first curated a random sample of sentences from 500 emails to be labelled under seven classes, with the additional class "None" represenitng that no given quality was present. This labelled set was then used to train the T5 and BERT models such that an individual set of words from any given email could have its most likely prevalent qualities identified by analyzing logit score outputs from the models implemented. The overall assessment of the sentences and sets of words in the emails was then used to build a combined probability score that the email has a "spike" in one or more of the six cognitive triaging qualities - indicating a likely attempt to phish. Combined with assessments from the other branches the intent is to supply a user with all the prospective information they may need to make a rational and informed assessment themselves given what emotions are present and what they are trying to elicit in the context of the email - information generated with the help of this branch.

Intent Analysis. The results have shown that the most promising models to implement this branch have been T5 and BERT, with T5 edging out BERT slightly with regards to accuracy and loss. To categorize common phishing intents that were present in emails, only short and apt descriptions have been added as tags to sets of word - such as "click link" or "download file". We randomly selected 500 emails from our datasets and conducted tagging manually. We then trained T5 to recognize malicious intents in the sentences of any email, with the objective of "cherry-picking" phishing intent in new emails users should pay particular attention to if other indicators such as cognitive triaging or the summary are suspicious. Figure 2 shows examples of intent evocations. We can note that the generated, concise summary contains the triggers and intent evocations at a much higher density, as we have suspected given our description in Section 2. By informing the user of the threshold criteria by which the tags are generated and the prospective danger of there being too many, the ambition is to achieve our set out objective of keeping the user alert. Experimentation at a larger scale is the next planned step to optimize this branch further.

4 Concluding Remarks

In this work, we devised a human-centric notification mechanism that extracts prospective psychological triggers, possible malicious intent, and a representative summary from emails. We then present the above in a meaningful way to the user for better decision-making and to elevate their learning of continuously evolving phishing patterns. Further examination with user studies and objective experimental analysis at a larger scale will give more insight into the effectiveness of this methodology. A concertized metric of trigger and intent density as a fraction of the total email length can also be examined to determine the metrics' correlation with whether an email is "phishy" or not.

References

[1] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies* 82 (2015), 69–82.

[2] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. 2021. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17. https://doi.org/10.1145/3411764.3445574

[3] Abdul Basit, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. 2021. A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems* 76, 1 (2021), 139–154.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[6] Amber Van Der Heijden and Luca Allodi. 2019. Cognitive triaging of phishing attacks. In *28th USENIX Security Symposium (USENIX Security 19)*. 1309–1326.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).