THE UNIVERSITY of EDINBURGH
informatics

aiai | Artificial Intelligence and its Applications Institute

PhishCoder: Efficient Extraction of Contextual Information from Phishing Emails

Tarini Saka (University of Edinburgh)
Nadin Kokciyan (University of Edinburgh), Kami Vaniea (University of Waterloo)

TULiPS
Technology Usability Lab in Privacy and Security

CHAI Lab

What is Phishing?

Phishing is a type of cyber-attack in which targets are contacted by email, telephone or text message by someone posing as a legitimate institution to trick individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. This information is then used to harm the users, organizations and society at large.

- Financial loss
- Reputational loss
- Loss of safety

Example of a Phishing Email

Delete Not spam Move to Categorise

Reminder: [Recent News] [Statement Appointment] Information Updates - New Notice : We sent Summary of your bill has expire, Renews your membership Sunday, Jan 22, 2020 [Statement:- [FWD] Case ID: M68EXXSC

Subject line confusing

From address domain is wrong

T TeamNetflix <noreplyrthismailerid33@ieuwgfiuwegfi
owe.[redacted]
Wed 22/01/2020 02:51

[redacted]

NETFLIX

Automatic payment.

Hi Customer,

Your Auto payment cannot process.
Your subscription period will end on Wed, January 22, 2020.

Asking to click link

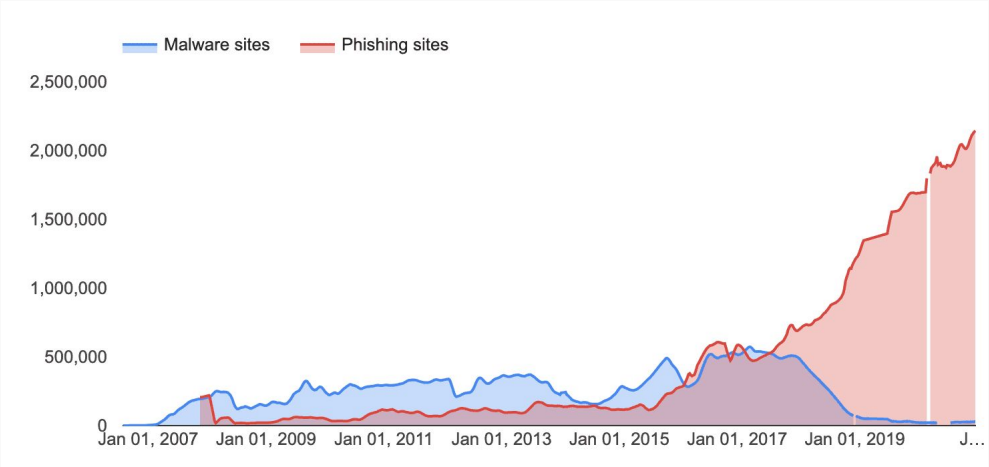
Click Here to update payment method

please update your payment methode for continue Netflix feature.

Grammatical Errors

The Netflix Team

Scale of the problem



Source: Tessian- Must-Know Phishing Statistics (https://www.tessian.com/blog/phishing-statistics-2020/)

1. APWG observed almost 5 MILLION phishing attacks in 2023, the worst year for phishing on record [2].
2. 94% of organizations were victims of phishing attacks [1].
3. 83% had multi-factor authentication (MFA) that was bypassed for the attack to succeed [1].

Source: [1] https://www.egress.com/blog/phishing/phishing-statistics-round-up
[2] https://apwg.org/trendsreports/

Organizational Phishing Defence

01 — Layer 1

Make it difficult for attackers to reach users.

02 — Layer 2

Help users identify and report suspected phishing emails.

03 — Layer 3

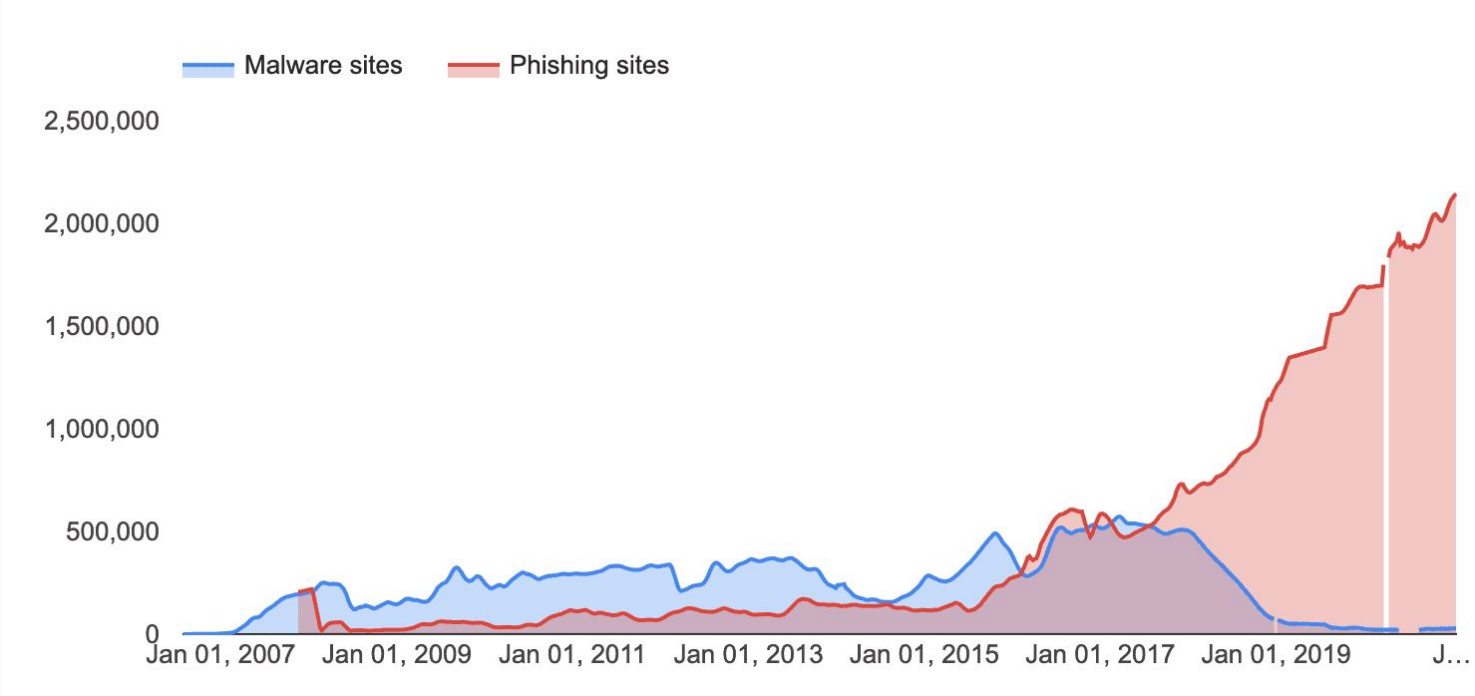
Protect against the effects of undetected phishing emails.

04 — Layer 4

Respond quickly to incidents.

Source: https://www.ncsc.gov.uk/guidance/phishing#section_3

Scale of the problem



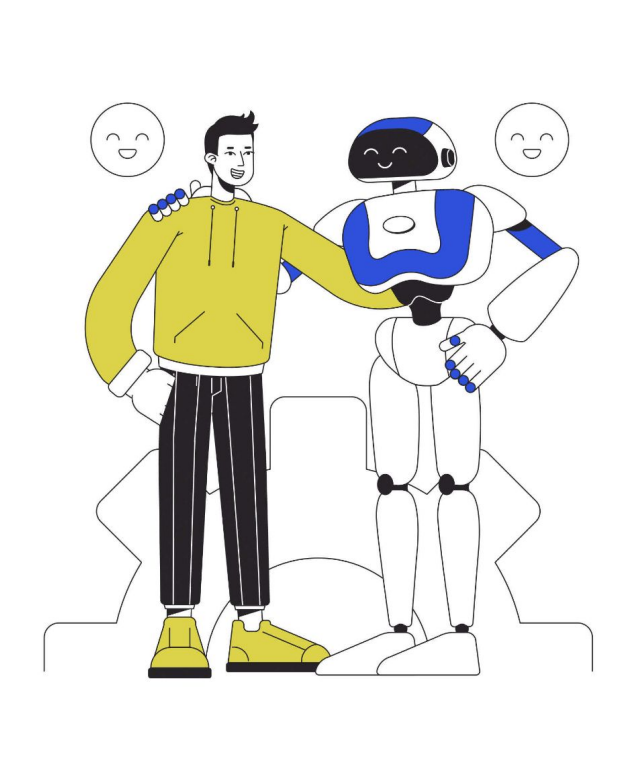
Source: Tessian- Must-Know Phishing Statistics (https://www.tessian.com/blog/phishing-statistics-2020/)

Solution: Automation and AI

1. An organization using AI-based security solutions can experience a reduction in costs associated with a data breach, from $6.71 million to $2.90 million.

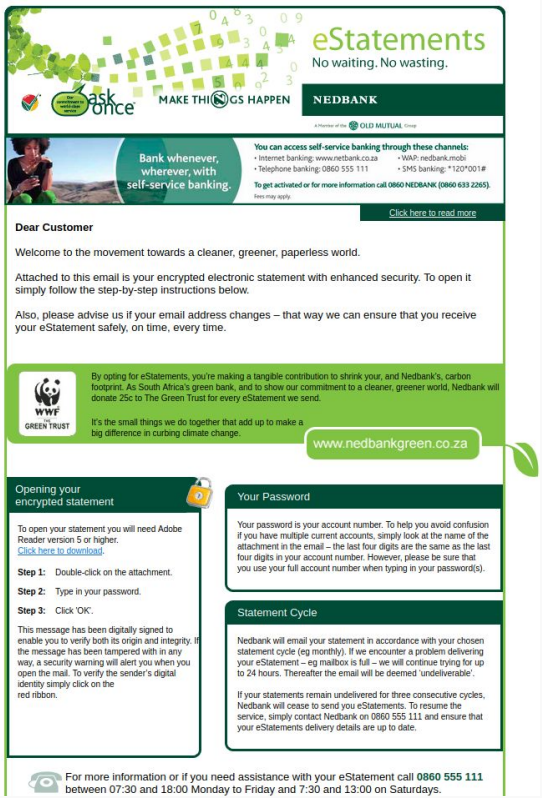
2. Security AI/automation was associated with a faster time to identify and contain the breach.

CHEAPER and FASTER!



Source: https://www.ibm.com/reports/data-breach

Research Gap: Email Variation



eStatements
No waiting. No wasting.

ask once
MAKE THINGS HAPPEN
NEDBANK
A Member of the OLD MUTUAL Group

Bank whenever, wherever, with self-service banking.

You can access self-service banking through these channels:
- Internet banking: www.netbank.co.za
- Telephone banking: 0860 555 111
- WAP: nedbank.mobi
- SMS banking: *120*001#

To get activated or for more information call 0860 NEDBANK (0860 633 2265).
Fees may apply.

Click here to read more

Dear Customer

Welcome to the movement towards a cleaner, greener, paperless world.

Attached to this email is your encrypted electronic statement with enhanced security. To open it simply follow the step-by-step instructions below.

Also, please advise us if your email address changes – that way we can ensure that you receive your eStatement safely, on time, every time.

WWF THE GREEN TRUST

By opting for eStatements, you're making a tangible contribution to shrink your, and Nedbank's, carbon footprint. As South Africa's green bank, and to show our commitment to a cleaner, greener world, Nedbank will donate 25c to The Green Trust for every eStatement we send.

It's the small things we do together that add up to make a big difference in curbing climate change.

www.nedbankgreen.co.za

Opening your encrypted statement

To open your statement you will need Adobe Reader version 5 or higher.
Click here to download.

Step 1: Double-click on the attachment.

Step 2: Type in your password.

Step 3: Click 'OK'.

This message has been digitally signed to enable you to verify both its origin and integrity. If the message has been tampered with in any way, a security warning will alert you when you open the mail. To verify the sender's digital identity simply click on the red ribbon.

Your Password

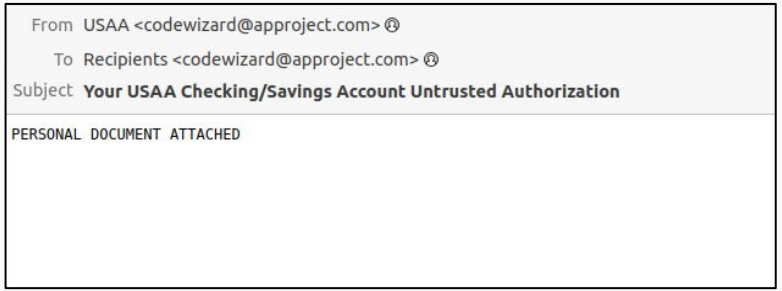
Your password is your account number. To help you avoid confusion if you have multiple current accounts, simply look at the name of the attachment in the email – the last four digits are the same as the last four digits in your account number. However, please be sure that you use your full account number when typing in your password(s).

Statement Cycle

Nedbank will email your statement in accordance with your chosen statement cycle (eg monthly). If we encounter a problem delivering your eStatement – eg mailbox is full – we will continue trying for up to 24 hours. Thereafter the email will be deemed 'undeliverable'.

If your statements remain undelivered for three consecutive cycles, Nedbank will cease to send you eStatements. To resume the service, simply contact Nedbank on 0860 555 111 and ensure that your eStatements delivery details are up to date.

For more information or if you need assistance with your eStatement call **0860 555 111** between 07:30 and 18:00 Monday to Friday and 7:30 and 13:00 on Saturdays.



From USAA <codewizard@approject.com> ☹
To Recipients <codewizard@approject.com> ☹
Subject **Your USAA Checking/Savings Account Untrusted Authorization**

PERSONAL DOCUMENT ATTACHED

Need for a CLEAR and CONCISE representation

Source: https://monkey.org/~jose/phishing/

Research Gap: Structured Context

----- Forwarded message -----
From: GOV UK Notify <danielnhs@pinkcontract.com>
To: "
Sent: Friday, 6 March 2020, 08:28:50 GMT
Subject: UK Updates on COVID-19

GOV.UK

The government has taken urgent steps to list coronavirus as a notifiable disease in law

As a precaution measure against COVID-19 in cooperation with National Insurance and National Health Services the government established new tax refund programme for dealing with the coronavirus outbreak in its action plan.

You are eligible to get a *tax refund* (*rebate*) of 128.34 GBP.

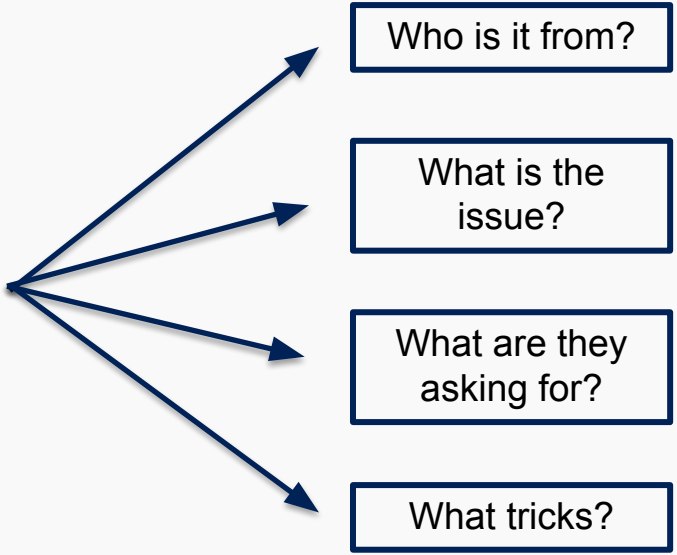
Access your funds now

The funds can be used to protect yourself against COVID-19(https://www.nhs.uk/conditions/coronavirus-covid-19/ precautionary measure against corona)

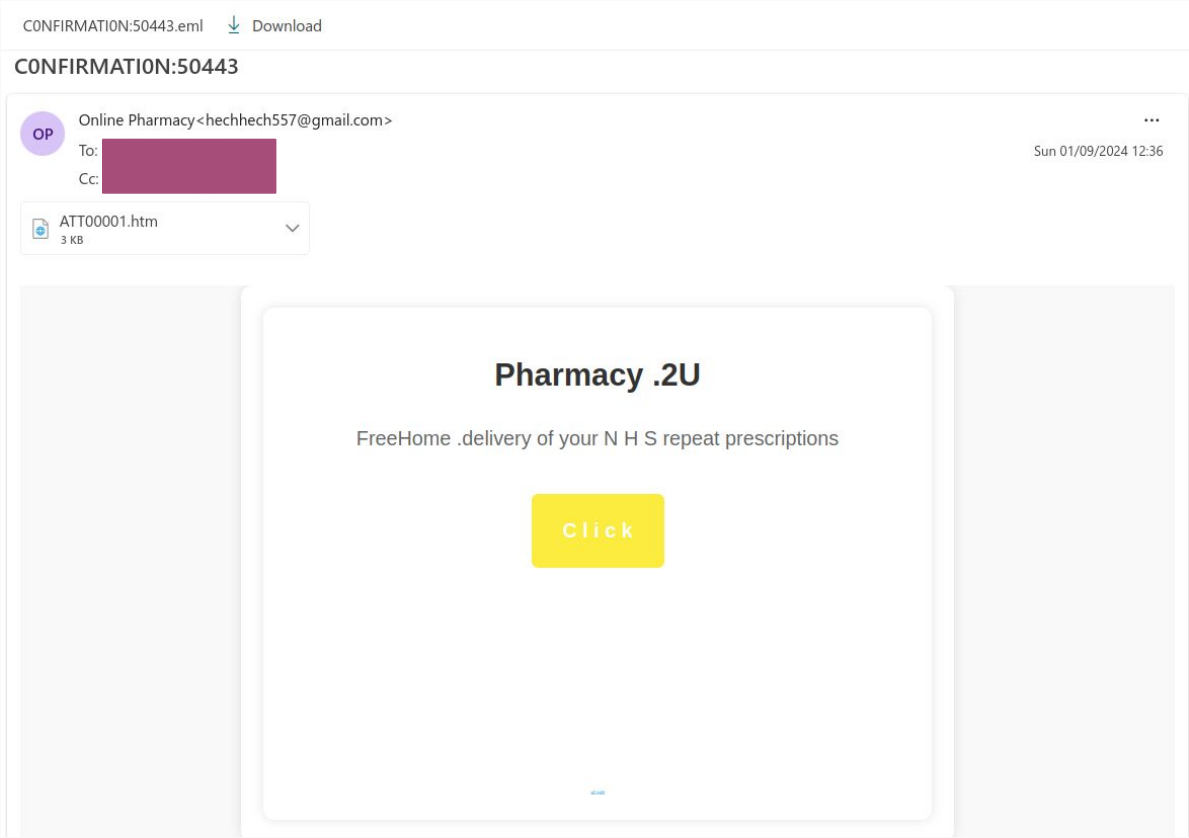
At 6.15pm on 5 March 2020, a statutory instrument was made into law that adds COVID-19 to the list of notifiable diseases and SARS-COV-2 to the list of notifiable causative agents.

From Government Gateway

This is an automatic email - please don't reply.

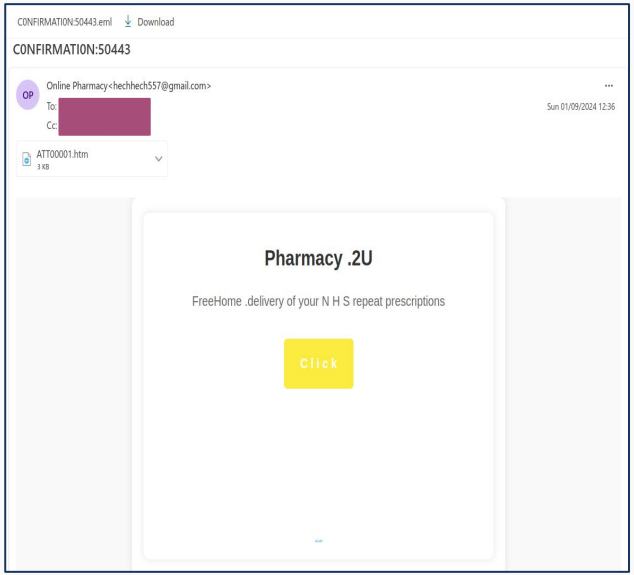


Research Gap: What Humans See?



Research Gap: What Computers See?

Research Gap: What is the correct context?



How to find the CORRECT context?

Past Research - A Phishing Codebook

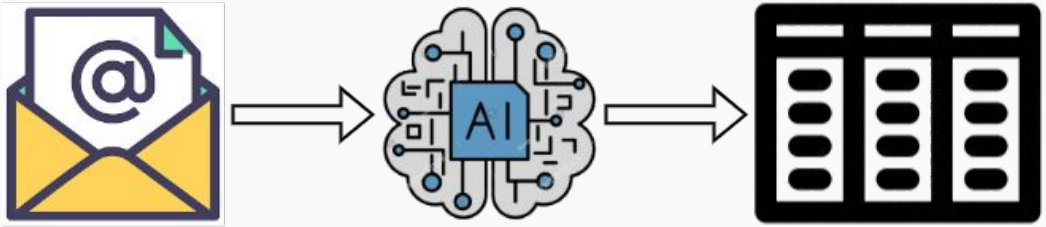
High-Level Code	Explanation	Sub-Codes
From- Company Name	Name of the organization being impersonated	in-vivo coding
From- Sector	Type of sector the email claims to be from	financial, email, document share, logistics, shopping, service provider, security, government, unknown
Salutation	Type of salutation used to address the recipient	name, email, generic, none
Threatening Language	Presence of threatening language	threat, none
Urgency Cues	Presence of time pressure or urgency cues	urgent, none
Action - Generic	The action being prompted in the email	click, download, reply, call, other, none
Main Topic	Main purpose of the email	in-vivo coding
Action - Specific	The reason provided to perform an action	in-vivo coding

Phishing Codebook: A Structured Framework for the Characterization of Phishing Emails - T. Saka, K. Vaniea, & N. Kökciyan; arXiv preprint

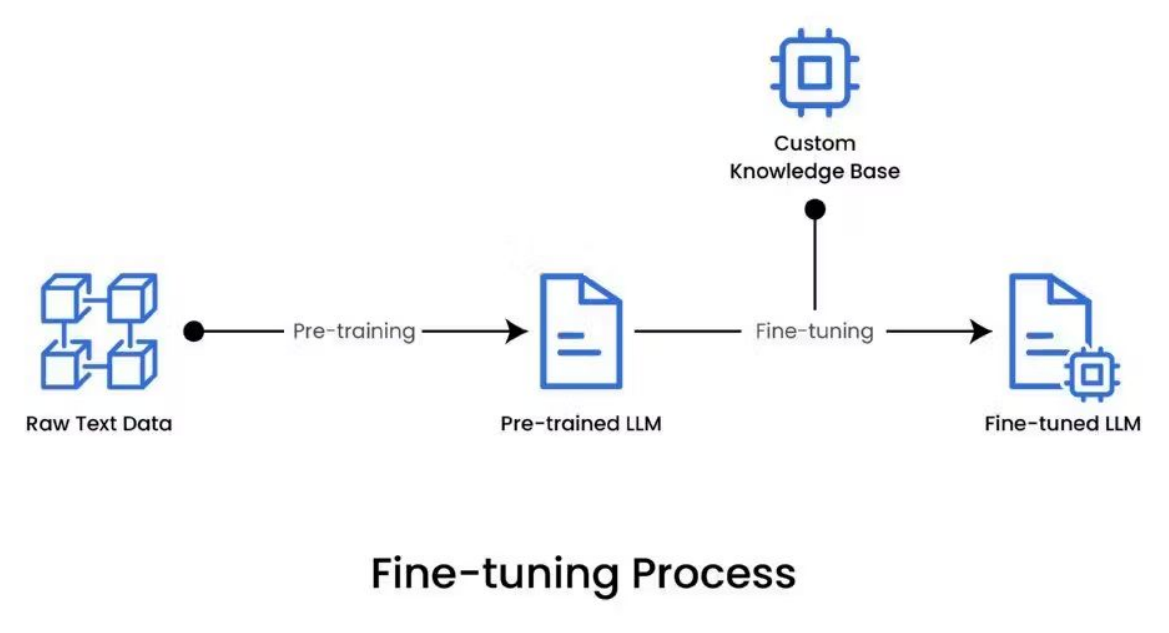
Code	Nedbank	USAA
From- Company	nedbank	usaa
From- Sector	financial	financial
Salutation	generic	none
Threatening Language	none	none
Urgency Cues	none	none
Action	download	download
Main Topic	encrypted electronic statement	personal document
Action Specific	attached to this email	document attached

PhishCoder

- In this paper, we introduce PhishCoder, a novel framework designed to extract contextual information from phishing emails.
- Now that we have a concise, contextual representation of phishing emails, we need a way to automate the process.
- This is a essentially information extraction from textual data - Language Models!
- Our focus is on human-centric features, which are often overlooked in traditional approaches, as they are the features that users notice when evaluating a potentially suspicious email.



Why Fine-tune Language Models?



1. Domain-specific expertise
2. Improved task performance
3. Customization
4. Efficiency
5. Faster convergence

Source: https://www.turing.com/resources/finetuning-large-language-models#what-is-fine-tuning,-and-why-do-you-need-it?

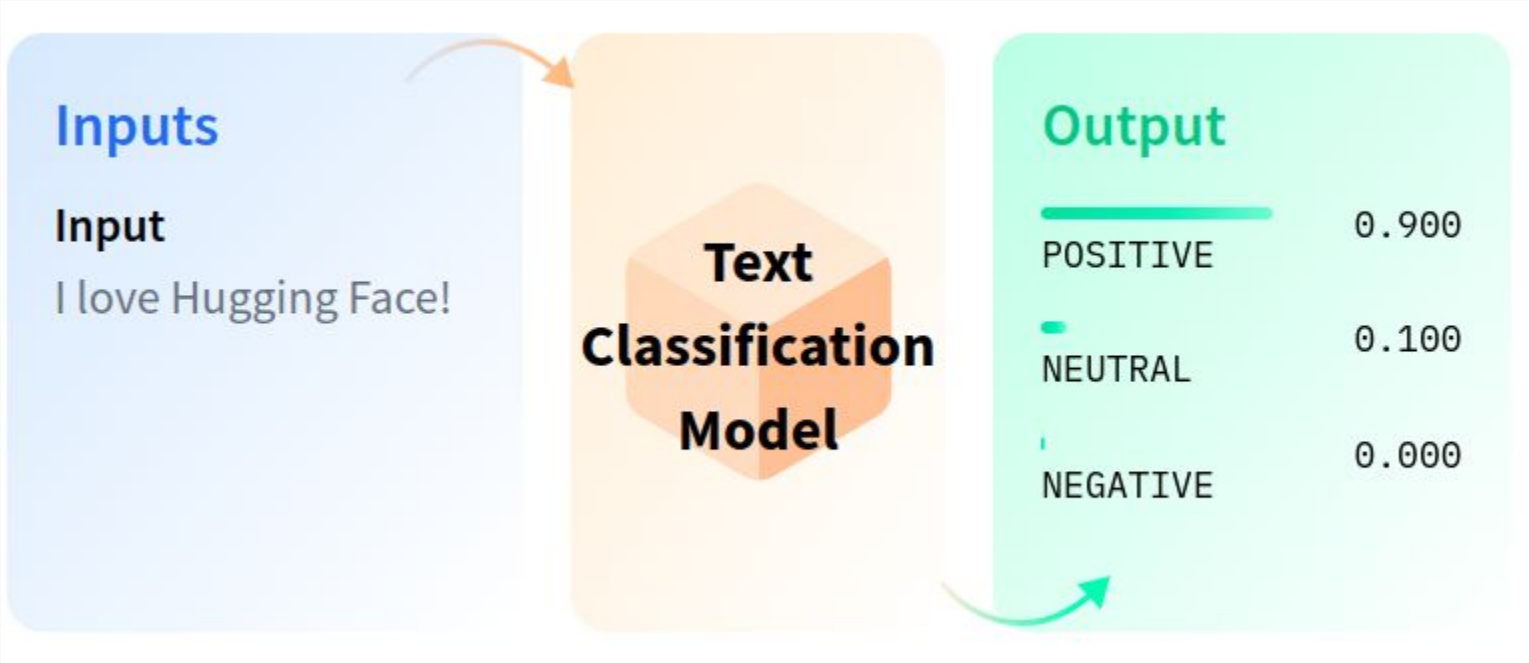
PhishCoder: Methodology

Step 1: Define the tasks for information extraction.

Task	Information Type	Explanation/Question	Labels
Text Classification Tasks			
TC1	From – Sector	Type of sector the email claims to be from	financial, email, document share, logistics, shopping, service provider, government, unknown
TC2	Action – Generic	The action being prompted in the email	click, download, other
TC3	Urgency Cues	Presence of time pressure or urgency cues	urgent, none
TC4	Threatening language	Presence of threatening language, tone	threat, none
Text Extraction Tasks			
TE1	From – Company Name	Name of the organization being impersonated	N/A
TE2	Action – Specific	The reason provided to perform an action	N/A
TE3	Main Topic	Main purpose of the email	N/A

Text Classification

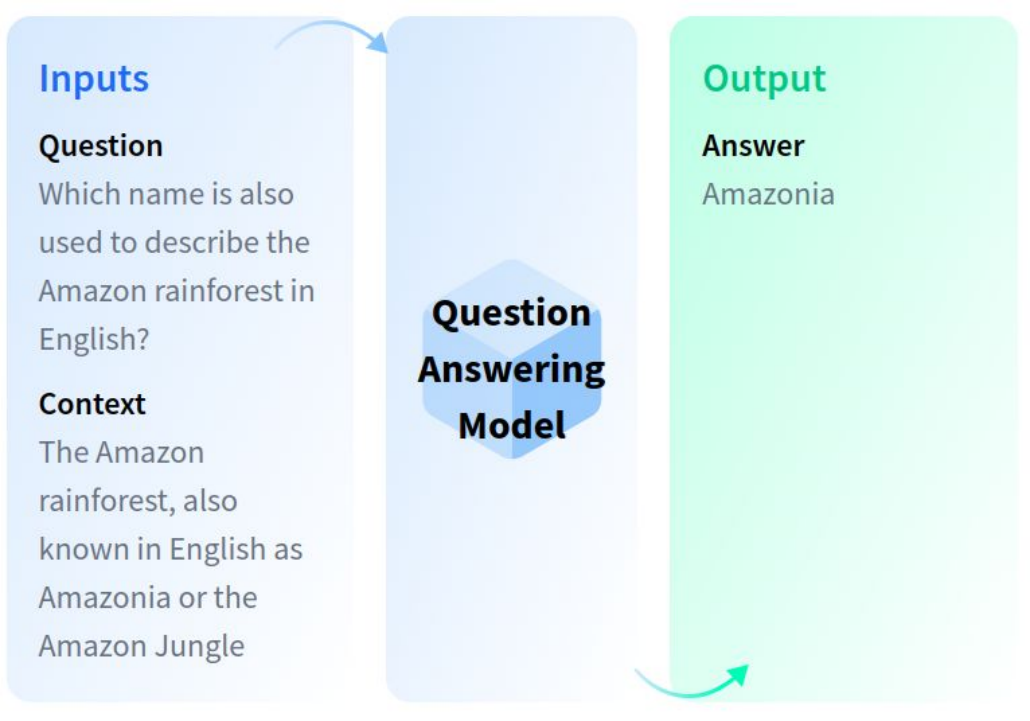
Text Classification is the task of assigning a label or class to a given text.



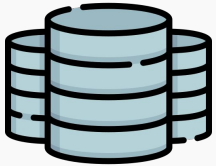
PhishCoder: Efficient Extraction of Contextual Information from Phishing Emails - T. Saka, K. Vaniea, & N. Kökciyan; SECAI 2024

Question Answering

Question Answering models retrieve the answer to a question from a given text: extractive or abstractive.



PhishCoder: Training Dataset



Nazario Phishing Dataset

- Publicly available hand-screened emails.
- 490 emails (D1)

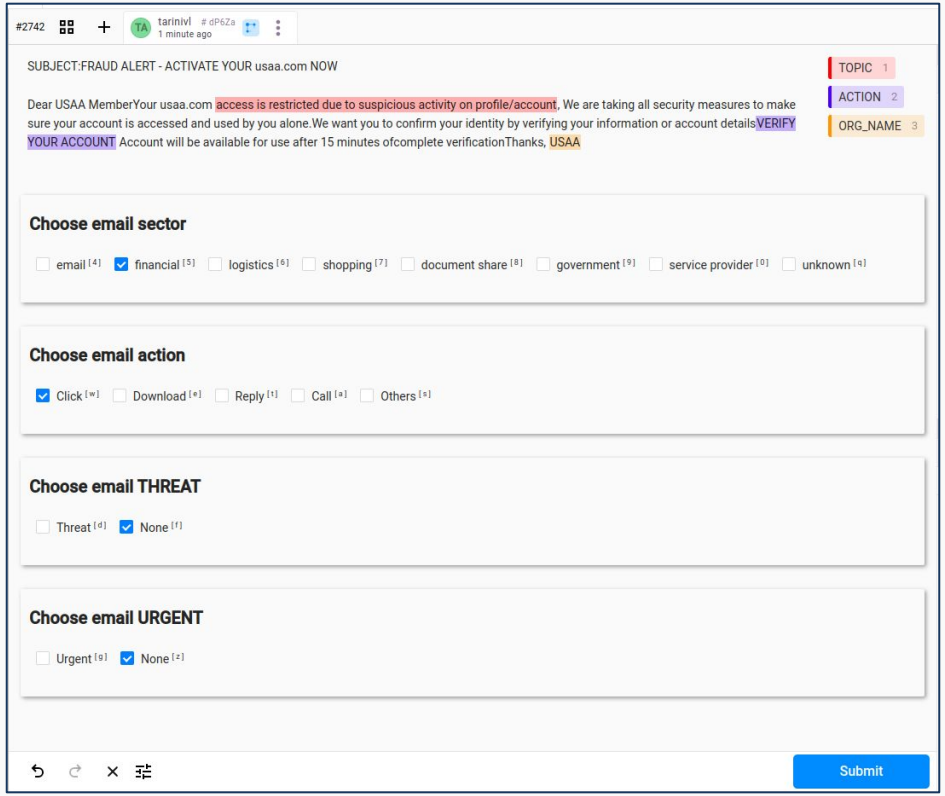
UoE Phishing Research

- Emails “donated” by staff to a research inbox.
- 31 emails (D2)

We created the final dataset of 521 emails by combining D1 and D2.

https://monkey.org/~jose/phishing/

PhishCoder: Labelling Data



https://labelstud.io/guide

PhishCoder: Fine-tune Language Models

Models

- **BERT:** Bidirectional Encoder Representations from Transformers.
- **RoBERTa:** Robustly optimized BERT.

Why?

- **Compact architecture**
- **Efficient in time and resources**
- **Simple to fine-tune**
- **Privacy-friendly (local processing)**

https://labelstud.io/guide

PhishCoder: Text Classification Results

Table 3: Evaluation of Action-Generic

Model Name	P	R	F1	Acc
bert-base	0.91	0.92	0.91	0.92
bert-large	0.93	0.94	**0.93**	**0.94**
roberta-base	0.91	0.92	0.92	0.92
roberta-large	0.93	0.94	0.93	**0.94**

Table 4: Evaluation of From Sector

Model Name	P	R	F1	Acc
bert-base	0.85	0.85	0.83	0.85
bert-large	0.88	0.87	0.87	0.87
roberta-base	0.96	0.94	0.94	0.94
roberta-large	0.98	0.96	**0.96**	**0.96**

Table 5: Evaluation of Threat Language

Model Name	P	R	F1	Acc
bert-base	0.83	0.83	0.83	0.83
bert-large	0.83	0.83	0.83	0.83
roberta-base	0.98	0.98	0.98	0.98
roberta-large	**1.00**	**1.00**	**1.00**	**1.00**

Table 6: Evaluation of Urgency Cues

Model Name	P	R	F1	Acc
bert-base	0.67	0.69	0.66	0.68
bert-large	0.93	0.93	**0.92**	**0.93**
roberta-base	0.87	0.87	0.87	0.87
roberta-large	0.89	0.89	0.89	0.89

https://labelstud.io/guide

PhishCoder: QnA Results

Table 7: Evaluation of Main Topic

Model Name	Exact Match	F1
bert-base	0.38	0.69
bert-large	0.30	0.68
roberta-base	**0.40**	**0.71**
roberta-large	0.32	0.69

Table 8: Evaluation of Action-Specific

Model Name	Exact Match	F1
bert-base	**0.58**	**0.78**
bert-large	0.47	0.77
roberta-base	0.49	0.74
roberta-large	0.43	0.70

Table 9: Evaluation of From - Company Name

Model Name	Exact Match	F1
bert-base	0.88	0.88
bert-large	**0.90**	**0.90**
roberta-base	0.88	0.88
roberta-large	0.85	0.87

PhishCoder: Results

- Our results show that fine-tuned language models are **promising** for extracting contextual information from phishing emails, offering a new direction for security research.
- Developed a multi-headed classification model for simultaneous task performance, but **individual models outperformed** it when trained with limited data.
- A major limitation is the issue of **class imbalance**, which can lead to biased models favoring the majority class.
- We need more annotated data from **different sources** to improve generalizability and performance.
- We got a new dataset from the Cambridge Cybercrime Centre.

PhishCoder: Research Contributions

RC1 → We introduce PhishCoder to capture the contextual nature of phishing emails by considering human-centric features.

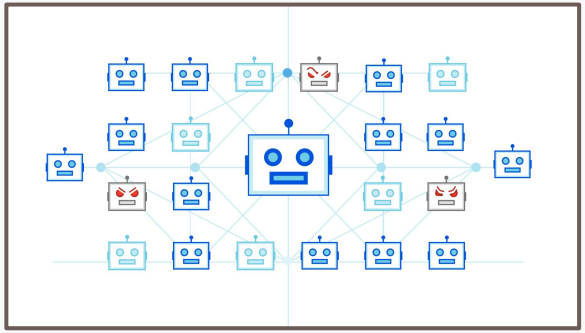
RC2 → Using real-world datasets and four pre-trained language models, we demonstrate their effectiveness in extracting contextual information from phishing emails.

RC3 → We explore the feasibility of a fine-tuned multi-task classifier to simultaneously perform the multiple tasks.

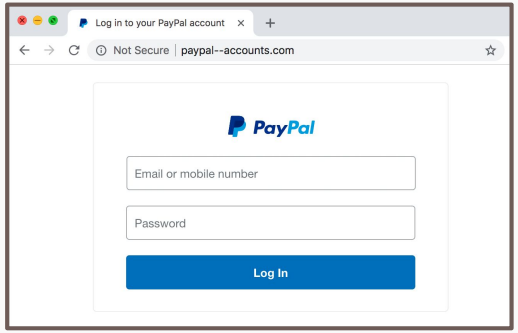
RC4 → We provide specific recommendations for using the PhishCoder outcomes.

Proposed Solution 1: Campaign Detection

A spam or phishing email campaign usually refers to a large number of emails sent by a common source. These emails share common characteristics such as the underlying fraud, the organization being impersonated, a malicious element, and the reaction it elicits.



Spam Botnets



Phishing Webpages

Research Goal - Identifying Campaigns

Bank.of .America. Bill Pay: Payee'(s) Added. Message.. - Mozilla Thunderbird

File Edit View Go Message Tools Help

Get Messages Write Tag

Reply Reply All Forward Archive Junk Delete More

From Bank of America Alerts <583-34493@telia.com>

To Recipients <583-34493@telia.com> 09/02/2017, 22:33

Subject **Bank.of .America. Bill Pay: Payee'(s) Added. Message..**

Dear Valid User,

A new payee has been added to your Bank. of. America. Bill Pay service.

The following payee (s) has been added to your Bank. of. America. list of payees in your Bill Pay service.

Please kindly Click On the <Log in bill Payment. System.>

for security reasons and fill out all the requested information,
for confirmation complete the verification process to prevent your account from being suspended.

==========
Internet Banking Account Alert.
Copyright © Bank of America, N. A. Member FDIC. Equal Housing Lender
powered by: Bank of America+

http://800donotcall.com/images/BankOfAmericalgr/

Wells Fargo Bill Pay: Payee'(s) Added. - Mozilla Thunderbird

File Edit View Go Message Tools Help

Get Messages Write Tag

Reply Reply All Forward Archive Junk Delete More

From Wells Fargo. Online Alerts. <gerdon@mwt.net>

To Recipients <gerdon@mwt.net> 29/01/2017, 12:56

Reply to wellsfargoonline@mail.com

Subject **Wells Fargo Bill Pay: Payee'(s) Added.**

A new payee has been added to your Bill Pay service

The following payee(s) has been added to your list of payees in your Bill Pay service.
Please Click on your mail below to download the attached file and fill out all the
requested information to

complete the verification process to prevent your account from being blocked.

Sincerely,
Online Services Team.
==========
Thank you for helping us keep your Wells Fargo account safe.
© 2017 Wells Fargo Corporation.Company. All rights reserved

powered by: Wells Fargo+

1 attachment: Action needed _Notifications.html 45.3 kB Save

Proposed Solution 2: User Assistance

AI-powered phishing-advice tool analyzes reported emails, providing tailored advice to users based on contextual phishing indicators, aiding decision-making understanding.

Source: PhishED: Automated contextual feedback for reported Phishing, In Proceedings of the Symposium on Usable Privacy and Security Poster Track. 2022.

Research Goal - Creating User Guidance

Get Messages | Write | Tag

Reply | Reply All | Forward | Archive | Junk | Delete | More

From Support <service@int-app.com>

To jose@monkey.org 24/02/2017, 18:30

Subject **[NETFLIX] Statement account be is available log in to re-active**

NETFLIX

Dear Valued Netflix User

Sorry for the interruption, but we are having trouble authorizing your Payment Method.
Please visit the account payment page at https://www.netflix.com/YourAccountPayment to enter your payment information again or to use a different payment method.
When you have finished, we will try to verify your account again.
If it still does not work, you will want to contact your credit card company.
To protect the informations of our customers, our system has temporarily placed restrictions on your account until your informations has been validated against our system.
You can validate your informations by either clicking on the link above or below, this will only take a few minutes and your account functions will be fully restored.

If you have any questions, we are happy to help. Simply call us at 0800-917812.
-The Netflix Team

LOGIN

Netflix Inc. : Netflix Corporate Headquarters 100 Winchester Circle Los Gatos, CA 95032.
You can un-subscribe to security alerts by configuring your online account.
We are sending this email to provide support for your personal online Netflix account.

AI Email Analysis

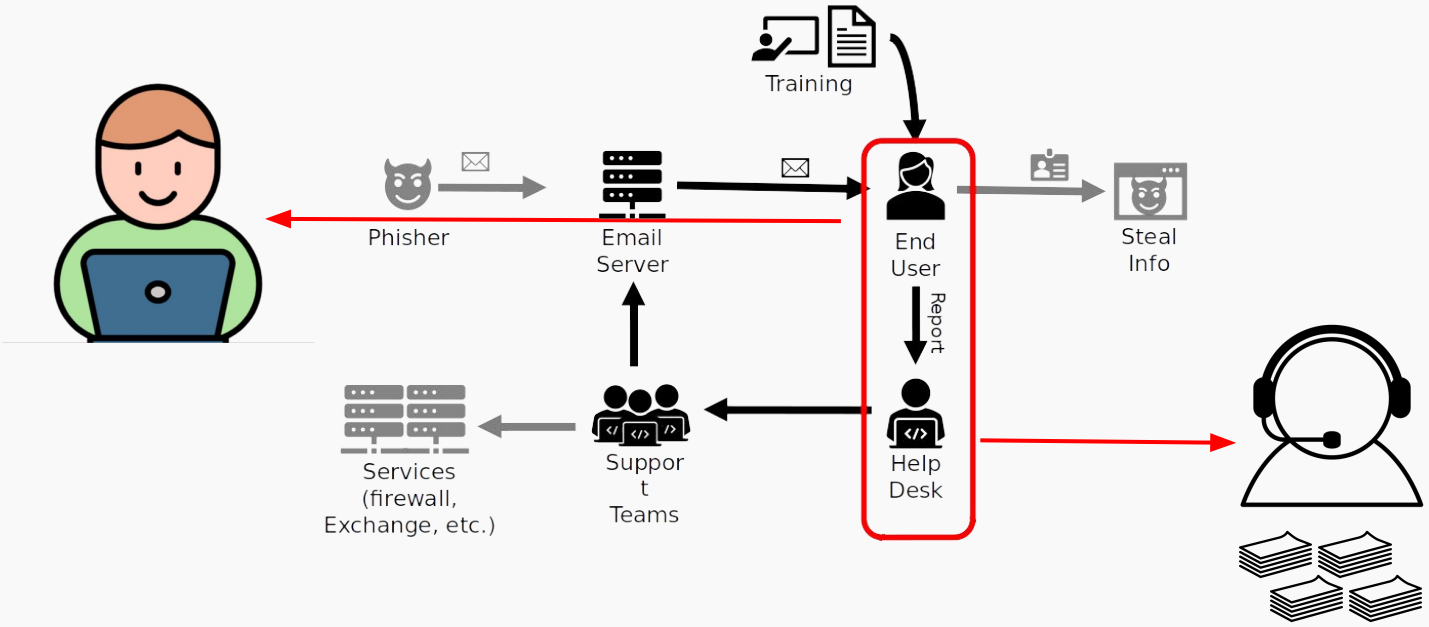
The following information was extracted from the emails using a AI model and based on this it is **95%** likely:

SCAM EMAIL

Organization Name	Action Requested
Netflix	Click

Sender domain	URL domain
'int-app.com' does not match Netflix domain	'srochnozaimy.com' does not match Netflix domain.

Research Aim - Efficient Phishing Mitigation



THE UNIVERSITY of EDINBURGH
informatics

aiai | Artificial Intelligence and its Applications Institute

Thank you :)

For any questions, comments or suggestions, please contact: Tarini Saka (tarini.saka@ed.ac.uk)

TULiPS
Technology Usability Lab in Privacy and Security

CHAI Lab