black hat MIDDLE EAST AND AFRICA

26 - 28 NOVEMBER 2024
RIYADH, SAUDI ARABIA

Harnessing Large Language Models for Detecting Malicious Attachments

Abhishek Singh, Kalpesh Mantri | InceptionCyber.ai

ORGANISED BY: Riyadh Alotaibi

tahaluf an informa company تحالف

الاتحاد السعودي للأمن السيبراني والبرمجة والدرونز
SAUDI FEDERATION FOR CYBERSECURITY, PROGRAMMING & DRONES

black hat

Disclaimer

Please note these slides represent the earlier version of NACE 1.0 which was presented at Black HAT MEA 2024 and AVAR 2024.

It does not accurately represent the current version of NACE.

black hat
MIDDLE EAST AND AFRICA

Abhishek Singh

- CTO and Founder of InceptionCyber.ai
- Led Research and Engineering at Cisco, FireEye, Microsoft
- Holds 40+ patents in cyber security, generative and predictive AI
- Authored 2 books on information security
- 2019 Reboot leadership award (Innovation category) SC Media, nominee for Peter Szor award
- Double MS in Computer Science & Information Security, Georgia Tech
- B.Tech in EE from IIT-BHU
- Post-graduate certificate in AI from IIT Guwahati

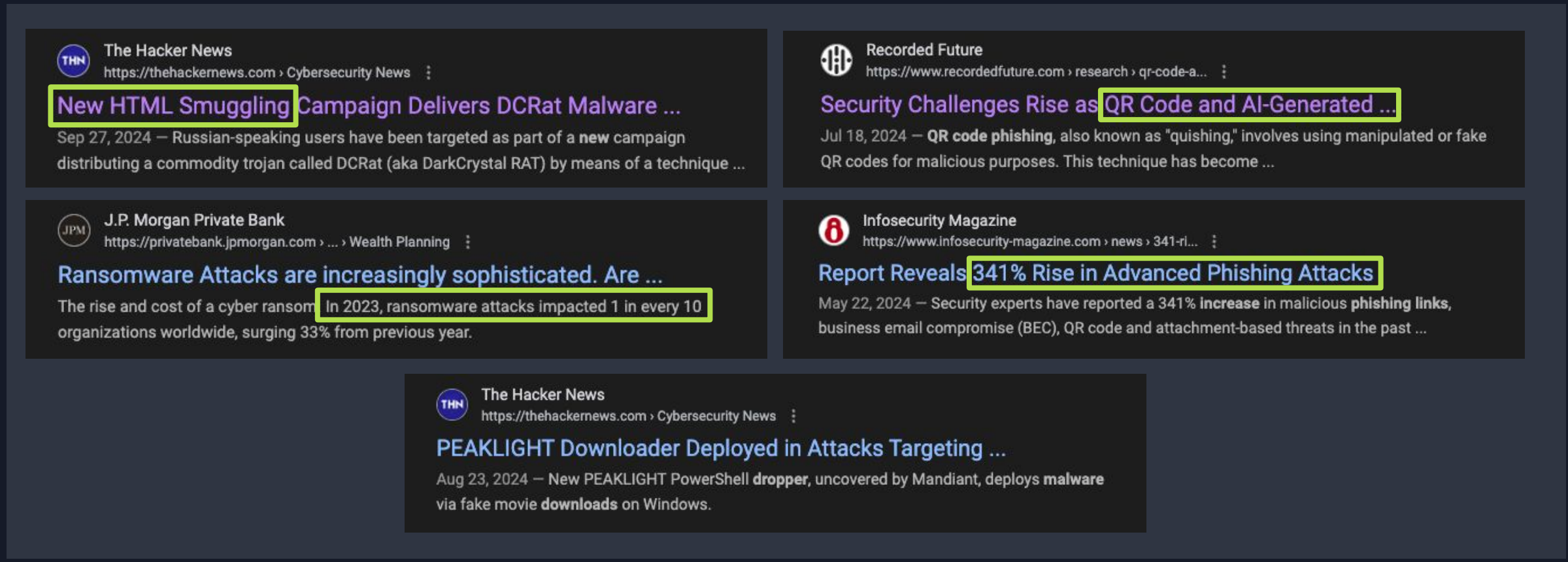
LinkedIn https://www.linkedin.com/in/abhisheksingh1/

Kalpesh Mantri

- Founding Principal Security Research Engineer at InceptionCyber.ai
- 12+ years of experience in Research and Engineering at McAfee, Quick Heal, Cisco
- Holds 3 patents in Design of Engine to Detect Malware and AI
- Presented research at Virus Bulletin, AVAR and CARO Workshop
- Led APT research and uncovered critical APT operations ‘Operation Side Copy’ and ‘Operation Honey Trap’ that target defense sectors
- Advance courses in AI from prestigious IIM Kozhikode

LinkedIn www.linkedin.com/in/kalpeshmantri

Current Threat landscape: Evasive Threats



Future Threat landscape: Generative AI for Attacks

BleepingComputer
https://www.bleepingcomputer.com › News › Security

OpenAI confirms threat actors use ChatGPT to write malware

Oct 12, 2024 — Although none of the cases described above give **threat actors** new capabilities in developing **malware**, they constitute proof that generative AI ...

SecureOps
https://secureops.com › blog › ai-attacks-fraudgpt

FraudGPT and WormGPT are AI-driven Tools that Help ...

Researchers have found ads posted on the Dark Web for an AI-driven hacker tool dubbed "**FraudGPT**," which is sold on a subscription basis and has been ...

http:\\www.hp.com
https://www.hp.com › press-releases › ai-generate-malware

HP Wolf Security Uncovers Evidence of Attackers Using AI ...

Sep 24, 2024 — Latest report points to **AI** use in creating **malware** scripts, **threat** actors relying on malvertising to spread rogue PDF tools, and **malware** embedded in image ...

BleepingComputer
https://www.bleepingcomputer.com › News › Security

Hackers deploy AI-written malware in targeted attacks

Sep 24, 2024 — Generative **AI** can help lower-level threat actors write **malware** in minutes and customize it for attacks targeting various regions and platforms (...

ts.blackhatmea.com/4-types-of-ai-threat-causing-global-disruption/

3. Automated malware aids antivirus evasion

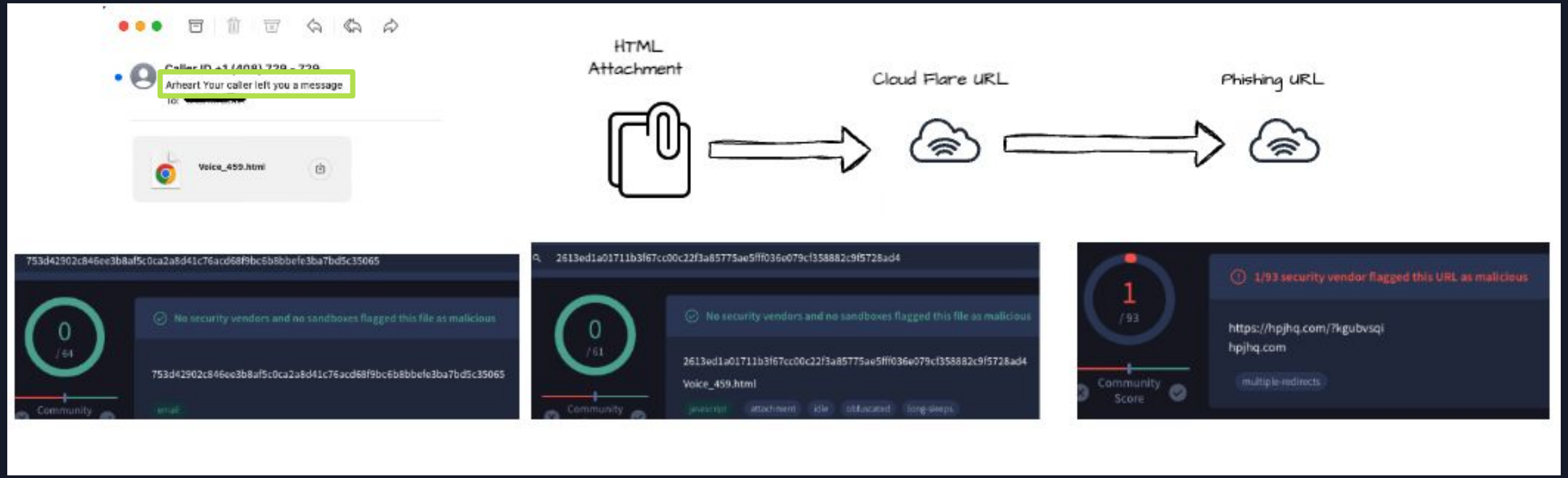
Threat actors are using AI to generate new malware variants very quickly. They use AI to analyse existing malware code and create slight variants – that are different enough to evade the signature-based detection models used by antivirus software.

Cyber criminals are also using AI to observe and analyse how malware reacts in a sandbox, and use this information to develop detection avoidance techniques in those environments.

Generative AI will be used to Learn, Adapt, and Craft Evasive Malicious Payloads at Unprecedented Scale

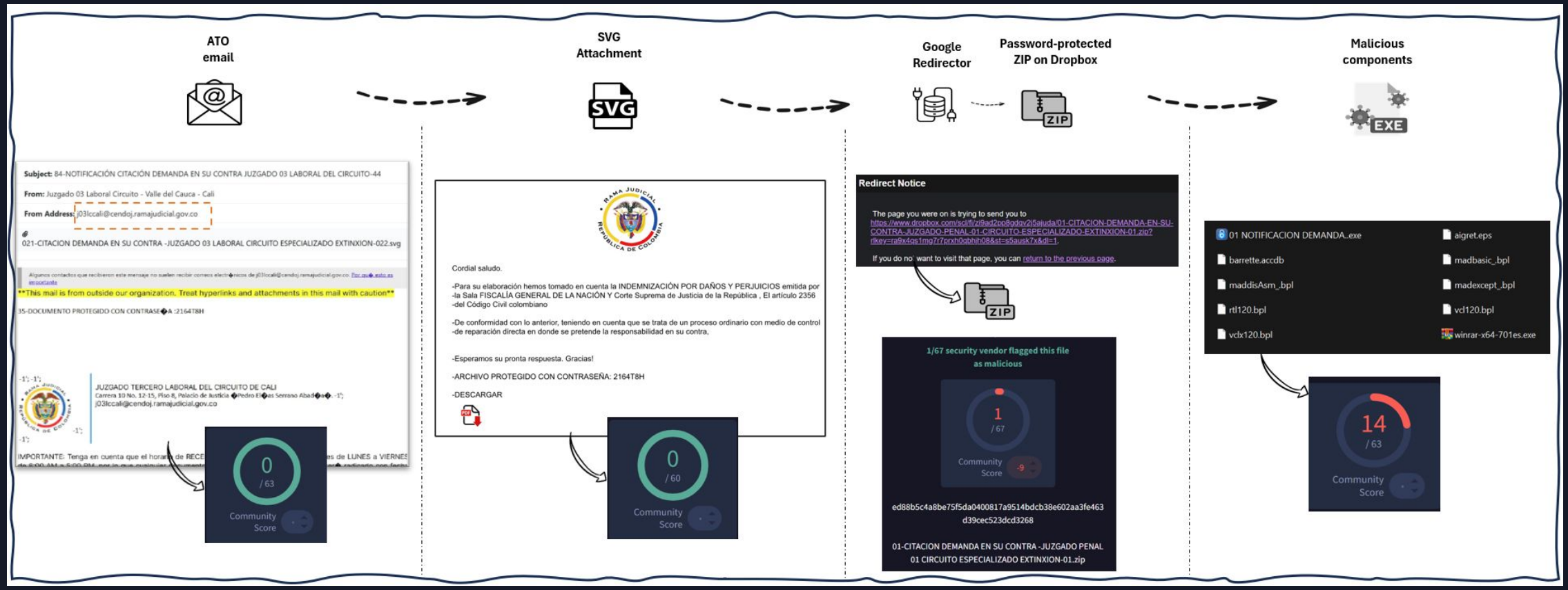
Understanding the Problem

Evasions (human or AI) hide malicious payloads in multi-stage attacks

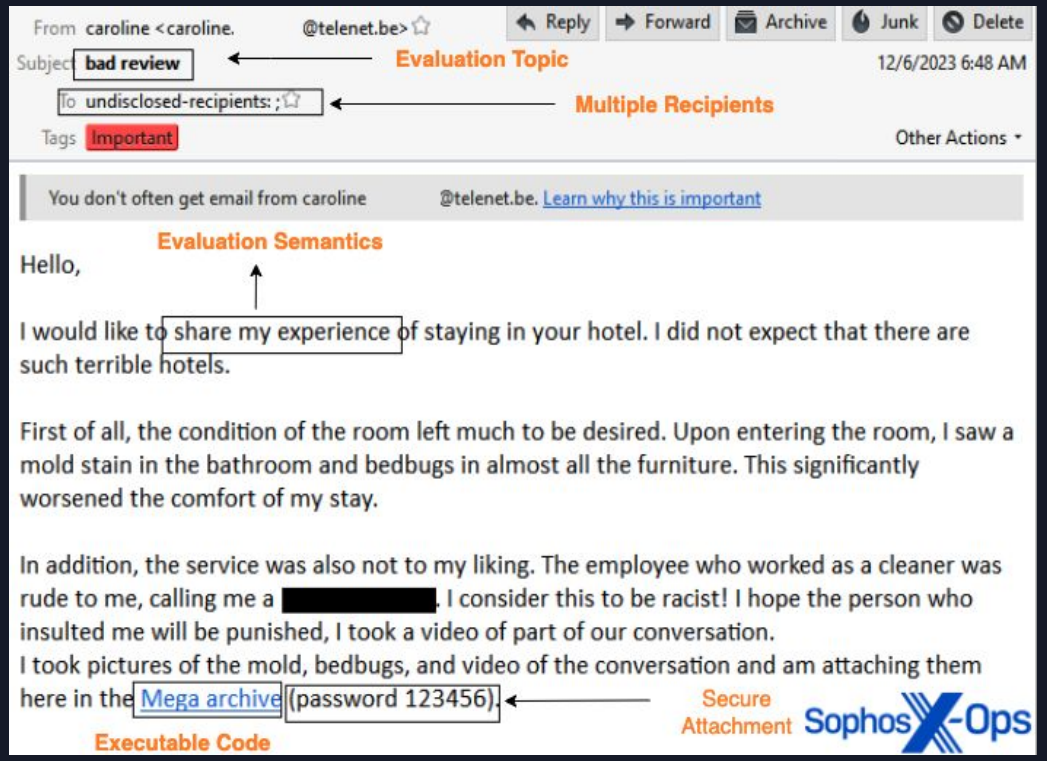


Understanding the Problem

Evasions (human or AI) hide malicious payloads in multi-stage attacks



Becoming Immune to Evasion: Solving From First Principles



Stopping threats has been focused on analyzing subsequent stages till malicious payload is seen.

Extract executable → detonate in sandbox → monitor behavior (Password Exfiltration)

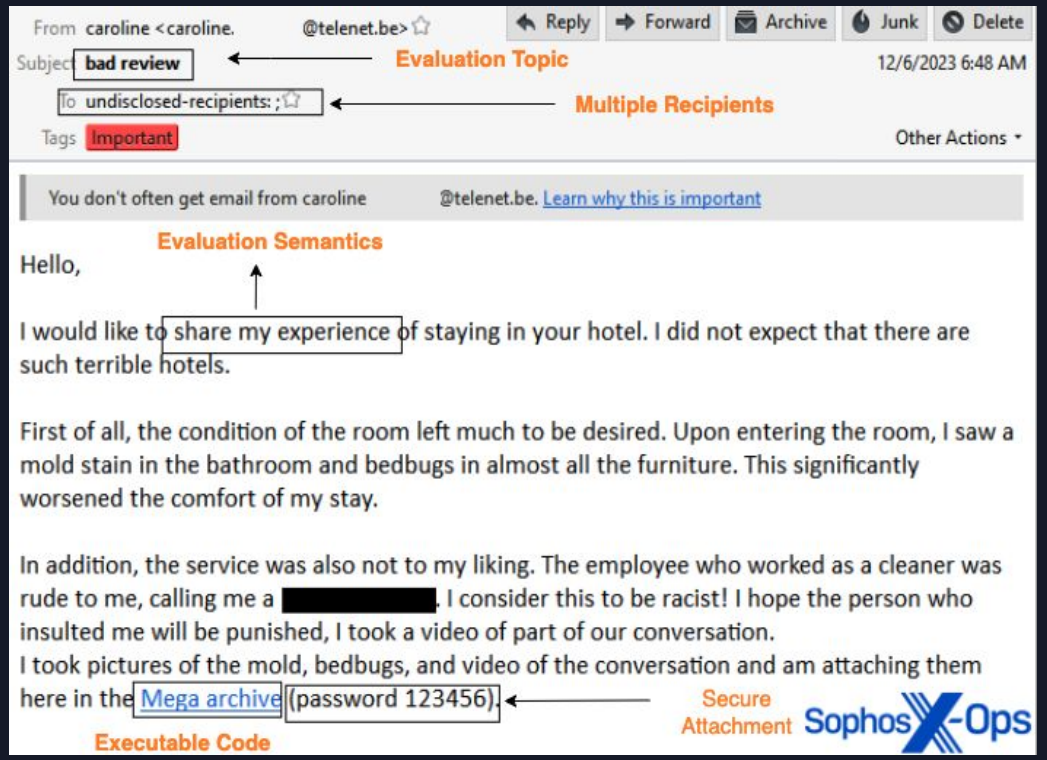
Therefore, evasions (human or AI) **Hide Malicious Payload** ⇒ Bypass Technology ⇒ Breach

Attack employed evasions (Signed files, large size, password protected, etc) → bypass sandboxes → Breach.

To change this paradigm, we must solve from first principles

A new methodology that doesn't require malicious payload/behavior ⟹ Immune to evasions ⇒ Inspection ⇒ Detection

Becoming Immune to Evasion: Intent-based Analysis



Derive Intent via Semantic and Thematic Analysis

Analysis

- Evaluation Communication having an executable code
- Sent to undisclosed recipients
- Sent from an external account

Verdict

- Unlikely behavior ⇒ Malicious Attachment

Leveraging **Semantic Analysis** as feature set removes reliance on Malicious Payload / fetching subsequent stages

⇒ Immune to Evasions

Design Steps : Semantics and Thematic Analysis for Classification

Step 1: Analyze Historic Threat Actor Emails

- Design a framework to extract semantic and thematic meaning from emails.

Step 2: Design an analysis system which does not need a malicious payload

- Examine emails to determine if they have semantic / thematic tactics used by threat actors
- Perform deep file parsing and analyze URLs
- Perform SMTP Header Analysis

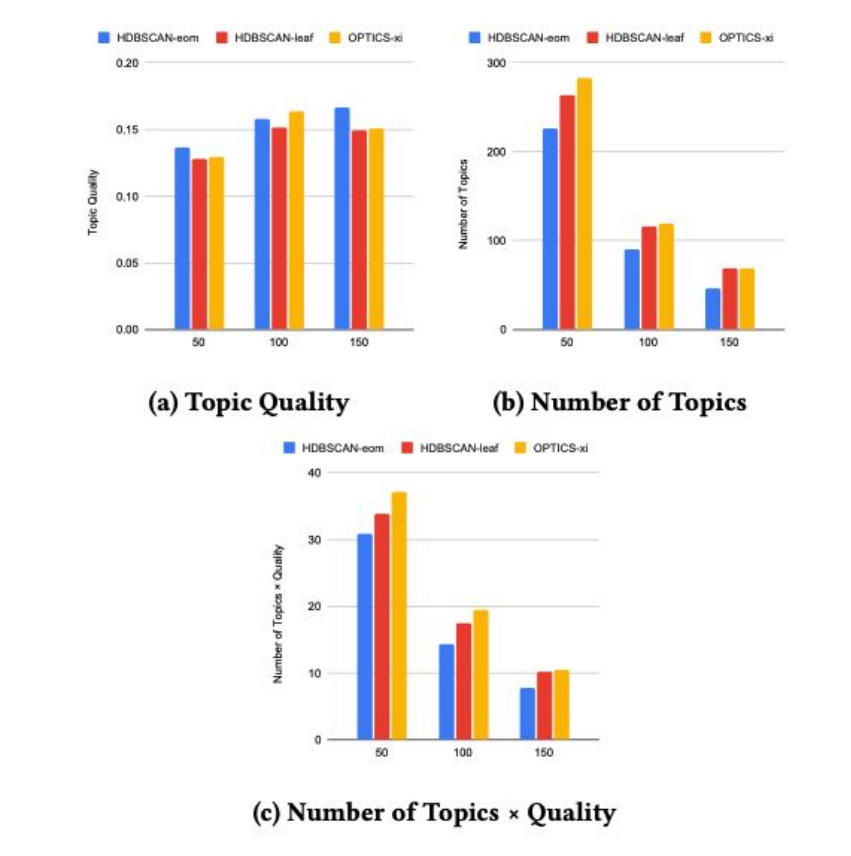
Leverage learnings from email semantics, deep file parsing and header analysis to classify attachments as malicious

Design Steps : Semantics and Thematic Analysis for Classification

Step 1: Analyze Historic Threat Actor Emails

- Design a framework to extract semantic and thematic meaning from emails.

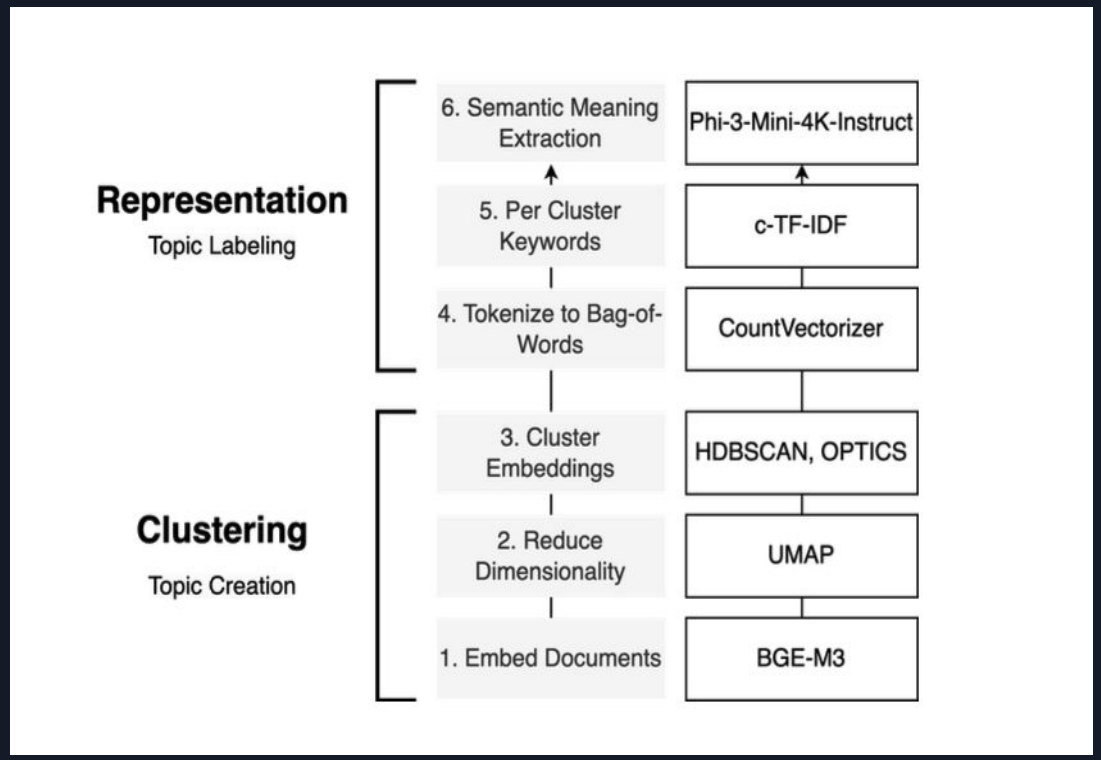
Becoming Immune to Evasion: Solving From First Principles



Experimentation with Unsupervised Clustering Algorithms

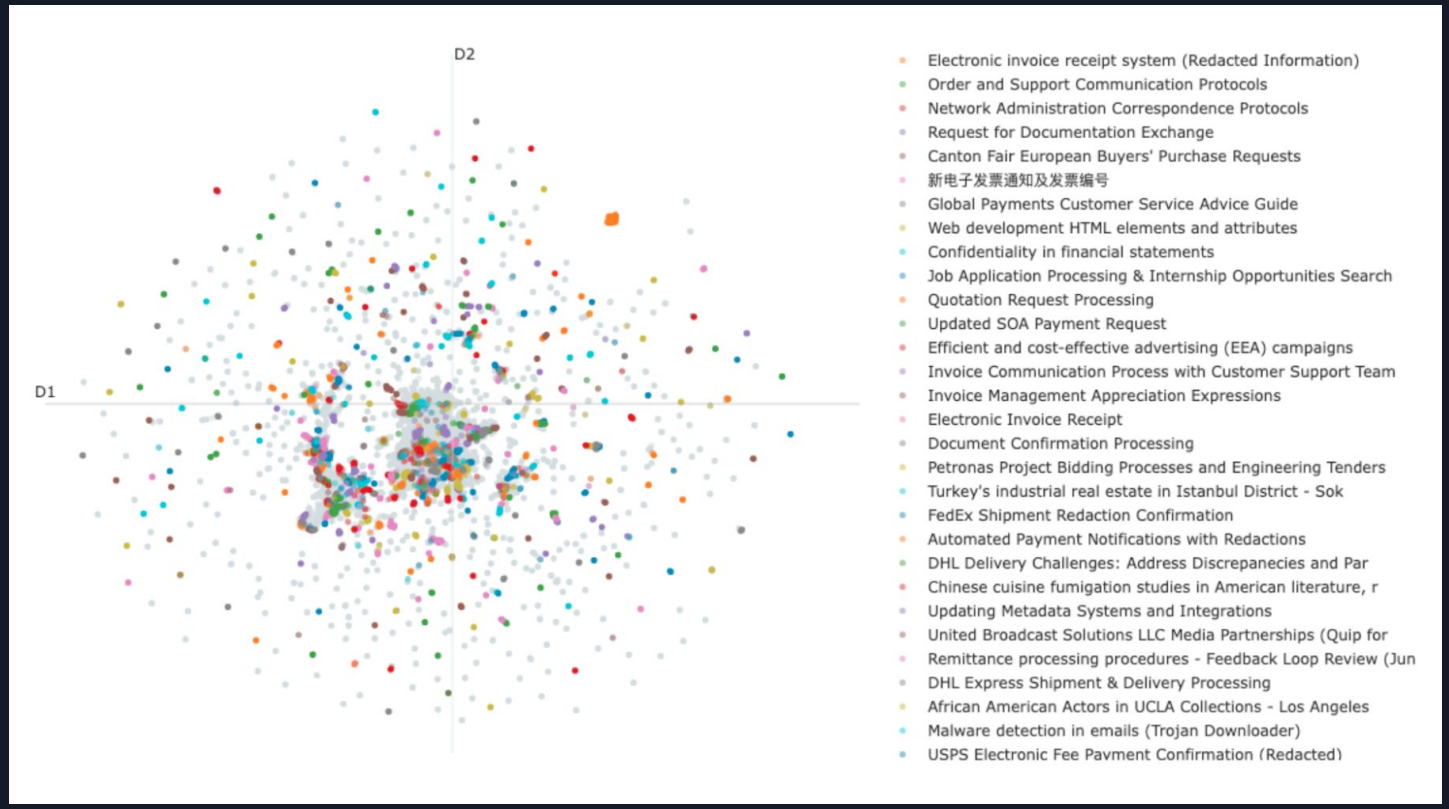
- Data Source:
  - Extracted email bodies, subject from historic emails used to deliver malware
- Algorithms Evaluated:
  - HDBSCAN - EOM
  - HDBSCAN - Leaf
  - OPTICS
- Evaluation Metrics:
  - Topic Quality - Measurement of granularity of cluster
  - Topic Coherence - Interpretability of a topic, closeness of words in topic
  - Topic Diversity - Unique words for all topics
- Result for Deciding OPTICS:
  - OPTICS produced 25% more topics while retaining 94.9% of the quality of HDBSCAN - EOM

Framework for extracting semantics from historic emails sent by threat actors to deliver malicious attachments



Name	c-TF-IDF Keyword Representation	Phi-3-Mini-4K-Instruct Semantic Meaning	Topic Hierarchy / Thematic Analysis
financial responding disapproval	[‘financial’, ‘responding’, ‘disapproval’, ‘very’, ‘topic’, ‘reporthello’, ‘monthly’]	Monthly Financial Response Evaluation Processing	‘financial’: [‘informational’]

Extracting semantics from clusters of historic malicious emails



Inferences

- Clusters denote topics which are getting repeated by threat actors to deliver malicious attachments and call to action URLs.

- **Extracted 1500+ semantics**
  Semantics which are extensively used by threat actors to deliver malicious attachments, URLs across languages.

  Details are in our arXiv:2407.08888 paper
  A. Yakymovych, A. SIngh et.al “Uncovering Topics and Semantics Utilized by threat actor to deliver Malicious Attachments”

Design Steps : Semantics and Thematic Analysis for Classification

Step 1: Analyze Historic Threat Actor Emails

- Design a framework to extract semantic and thematic meaning from emails.

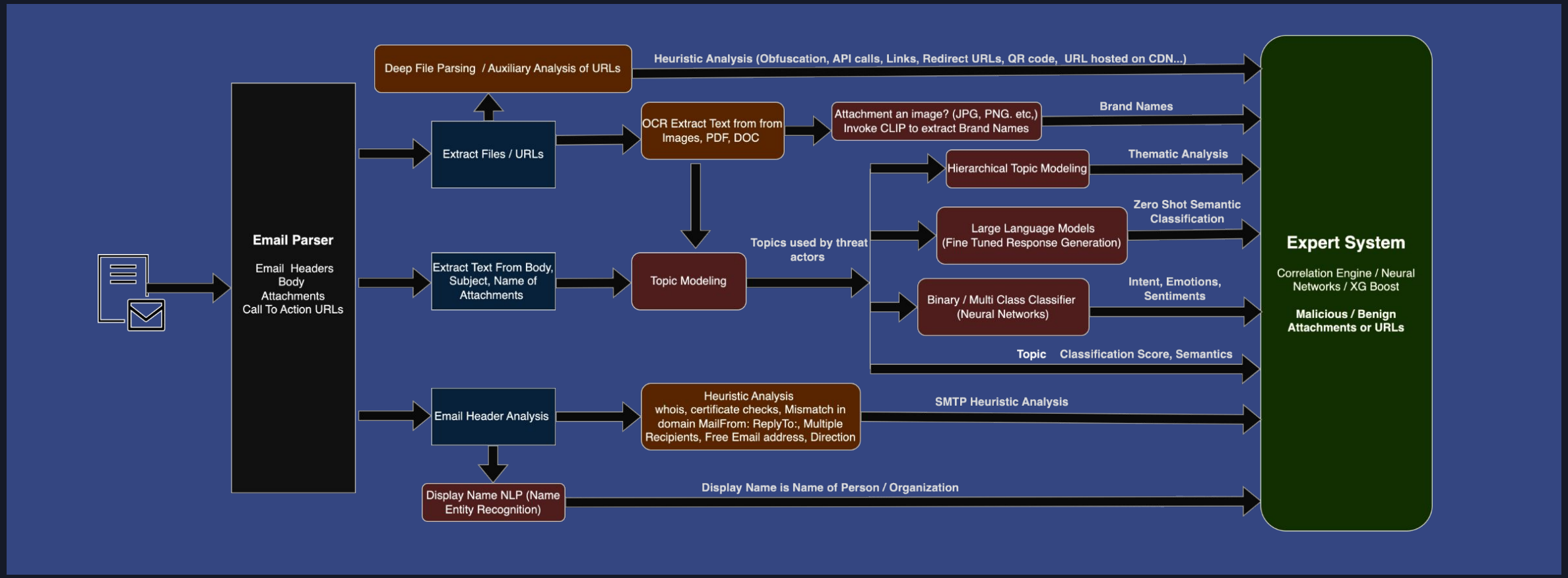
Step 2: Design an analysis system which does not need a malicious payload

- Examine emails to determine if they have semantic / thematic tactics used by threat actors
- Perform deep file parsing and analyze URLs
- Perform SMTP Header Analysis

Leverage combination of learnings from email semantics, file parsing results, and header analysis to classify attachments or URLs as malicious

Design of a Neural Analysis and Correlation Engine (NACE)

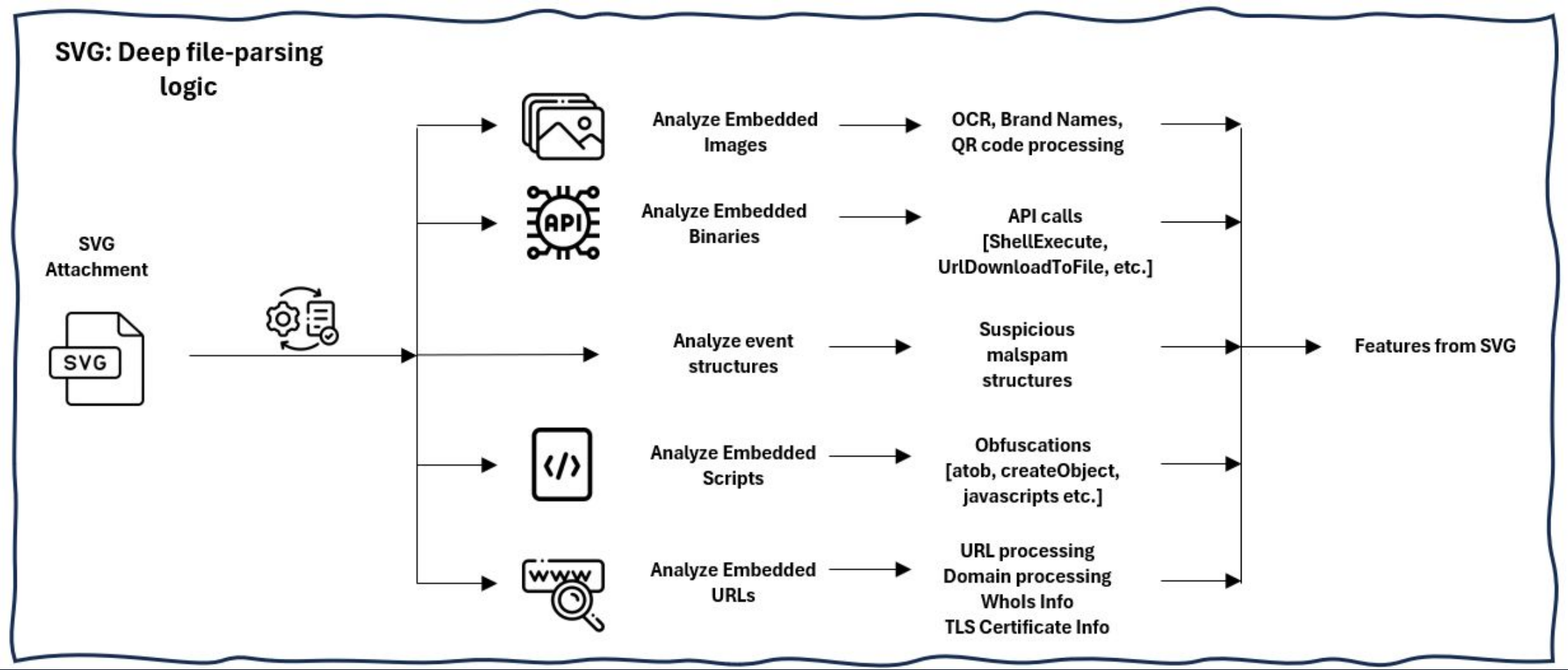
Leveraging Topics and Semantics to detect Malicious Attachments and URLs



Deep File Analysis of NACE

NACE performs deep file parsing, text extraction via OCR,, Brand Extraction using CLIP, API Invocation, Obfuscation, gathering auxiliary information (whois, certificates etc....) of any embedded URLs in a file for 20+ file formats:

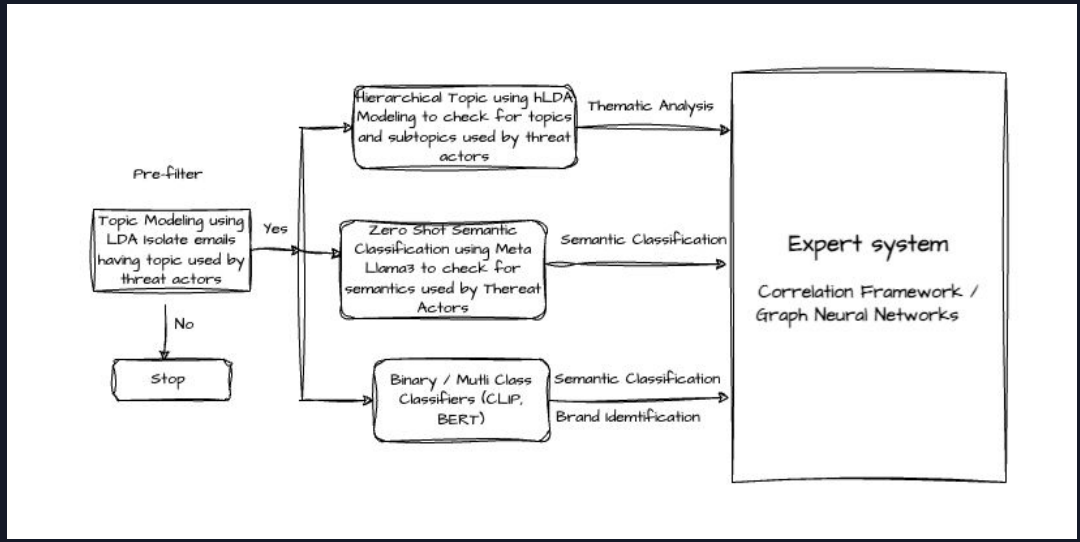
- Document File Formats (Office file types, PDF, OneNote, etc.)
- Archive File Formats (ZIP, RAR, ISO, ZPAQ, etc.)
- Image File Formats (PNG, JPEG)
- Script File Formats (VBS, JS, PY, etc.)
- Markup and Web File Formats (HTML, SVG, HTA, XML, etc.)
- Executables (EXE, LNK, VBE, BAT, etc.)


SVG: Deep file-parsing logic
SVG Attachment
SVG
API
Analyze Embedded Images
OCR, Brand Names, QR code processing
Analyze Embedded Binaries
API calls [ShellExecute, UrlDownloadToFile, etc.]
Analyze event structures
Suspicious malspam structures
Analyze Embedded Scripts
Obfuscations [atob, createObject, javascripts etc.]
Analyze Embedded URLs
URL processing
Domain processing
Whois Info
TLS Certificate Info
Features from SVG

Semantic & Thematic Analysis of NACE

Isolate Embedded Semantics & Thematic meaning in an Email

- **Topic Modeling prefilter to Invoke Semantic Analysis**
  - LDA: Excels in Identifying distinct topics . Suitable
  - BERT Topic: Excels in identifying semantic similarity and not distinct topics. Lack of fine tuned Topic result in FP in detections.
- **Hierarchical Topic Modeling : Topic & Subtopic in a text**
  - hLDA: Consistent results with fixed seed, fine-tuned Topic/Subtopics
  - HDP: Non-parametric Bayesian Approach, Random sampling, Inconsistent Results across multiple runs. Not suitable
- **Zero Shot Semantic Classification: Semantic embedded in an email**
  - Leverages prompt engineering for fine -tuned response generation
    - Fine tuning parameters (temp, top_p) passed to Large Language Model. (LLMs) restricted creativity mode.
    - Identified precise semantics embedded in text,
    - Immune to variations



Expert System for Decision Making

Correlation Engine

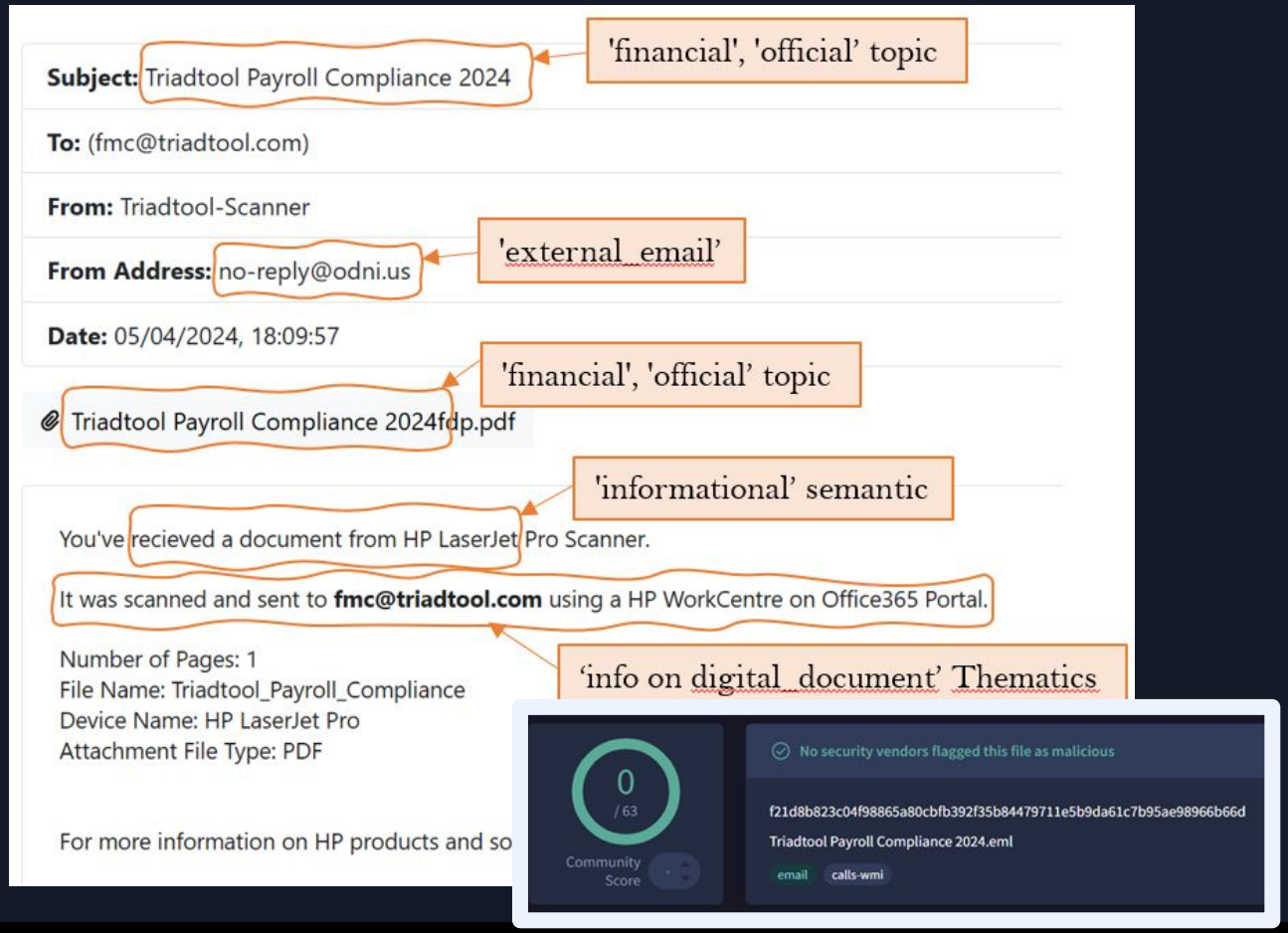
- Correlates Semantic analysis, Thematic Analysis, Topic Modeling, Deep File Parsing, SMTP Headers to decide if Attachment is malicious

Graph Neural Networks

- Nodes Semantic Analysis, Deep File Parsing and SMTP Headers to decide malicious or benign files / URLs

```
IF
(file_semantic contains any item that matches:
'has_executable_code_svg' OR
'has_executable_code_raw_parsing_svg')
AND
(body_semantic contains any item that matches:
' 'financial_semantic' OR
'p_invoice_c_financial' )
AND
(sender_semantic contains any item that equals:
'is_probable_external_email')
THEN
FLAG as potential threat
```

Case-Study:
Topics, Semantics, Thematic Analysis



Header Semantics

Sender Domain analysis:
- WhoIs info, Newly registered, Cloud name_servers etc.
- TLS certificates, Free email domain, Domain buckets etc.
- ...

Header:
- Anomalies identification (Sender is Non-Permitted IP, Return Address / MessageID Anomaly etc.)
- XMailer, UserAgent, Multi-Recipients identifications etc.
- Mail directions
- ...

Subject Topic Identification

Attachment-Name Meta

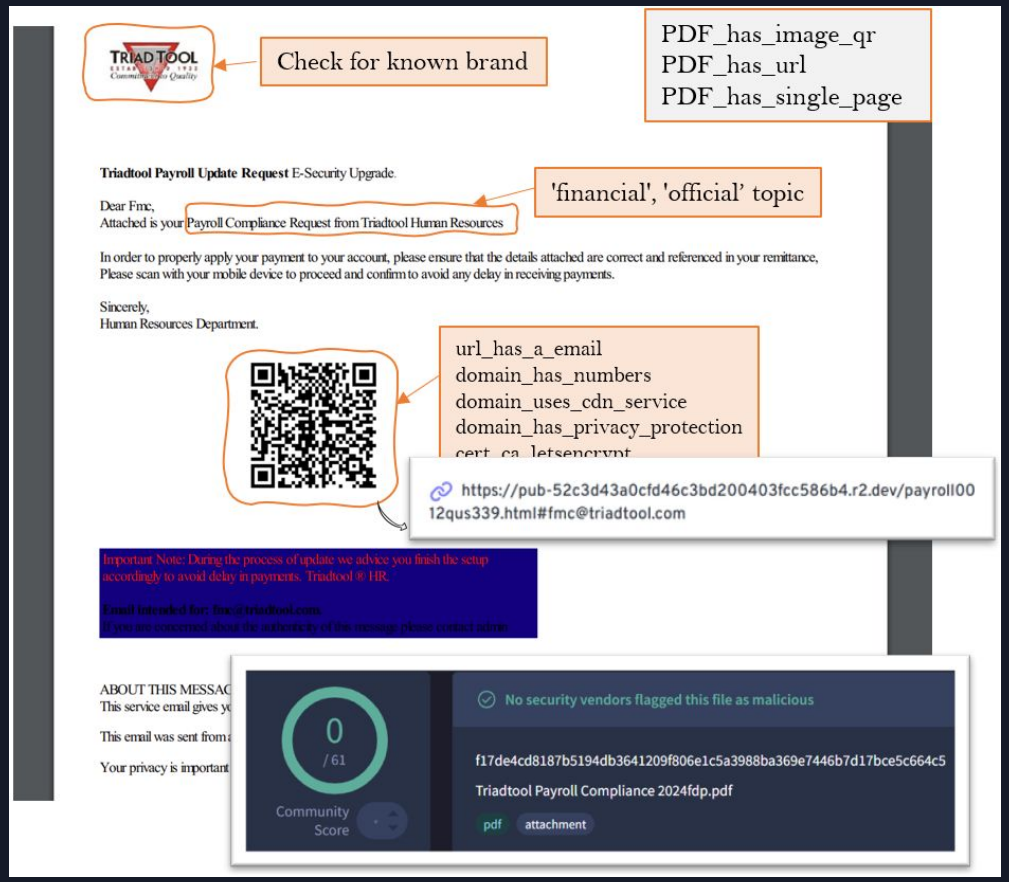
Subject Topic Identification Financial, Official

Body Semantics and Thematics

Topic modeling: Official, Financial
Thematics Analysis : Informational [Digital_Communication]
LLM Classification (Zero-Shot Semantics): Informational Semantics
URL extraction and analysis

Image analytics:
- Text extraction / OCR
- QR code analysis
- Brand Identification

Case-Study Continued:
Attachment Semantics, Image Processing, URL analysis



Attachment Semantics

PDF Structural Analysis
PDF Metadata identification
PDF Tags Counter
Image extraction
URL extraction
PDF Topic identification
...

Image Processing

OCR, Brand Identification, QR Code Identification, etc.

URL Analysis

Suspicious URL format analysis
Identification of suspicious domains
Whols information
Domain Creation info
Domain Categories
TLS Certificates analysis
Analyzing Document sent as URLs
...

Case-Study Continued: Expert System

Header

'is_probable_external_email'
'is_first_email'

Body

'informational semantic' semantics
'p_informational_c_digital_communication' thematics

Verdict

'QR code phishing detection'

Subject

'encoded_subject'
'financial','official' topics

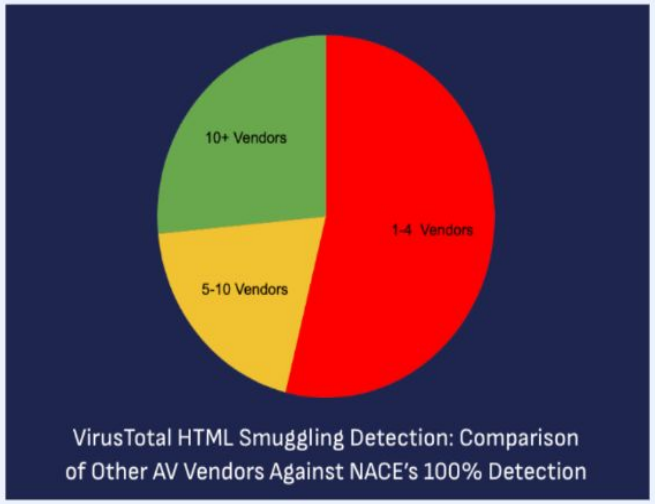
Attachment

'attachment count == 1',
'PDF has single page'
'PDF has image qr',
'PDF has qr code semantic'
'domain uses cdn service',
'cert_ca_letsencrypt'
...

An external email exhibits language patterns that suggest a call to action on a financial semantic on a PDF attachment, generated using digital communication technology. The attachment contains an embedded QR code hosted on a content delivery network (CDN) and secured with a certificate from a free provider. These factors collectively result in the email being classified as QR code phishing Scam.

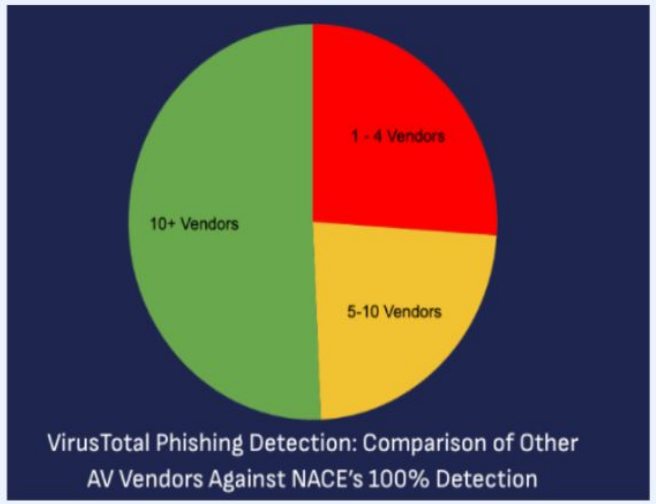
Benchmark against other Technologies

- Data Source: ~13K Samples, 2024 Evasive threats (HTML Smuggling, Phishing, Downloaders, Dropper, etc) from Viru Total
- VirusTotal: 70 AV Vendors, 70 URL Black List, 10 Sandboxes
- Results: 99% of coverage, ~44% of the evasive threats detected by NACE were missed by 95% of the AV Technologies



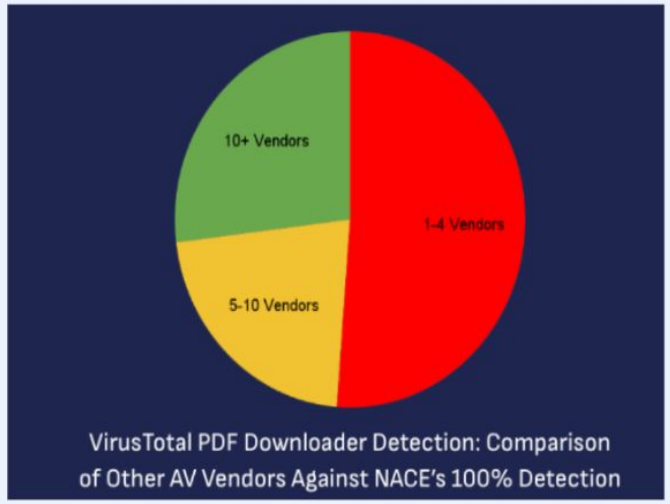
VirusTotal HTML Smuggling Detection: Comparison of Other AV Vendors Against NACE's 100% Detection

51% of HTML smuggling detected by NACE missed by ~95% of AV



VirusTotal Phishing Detection: Comparison of Other AV Vendors Against NACE's 100% Detection

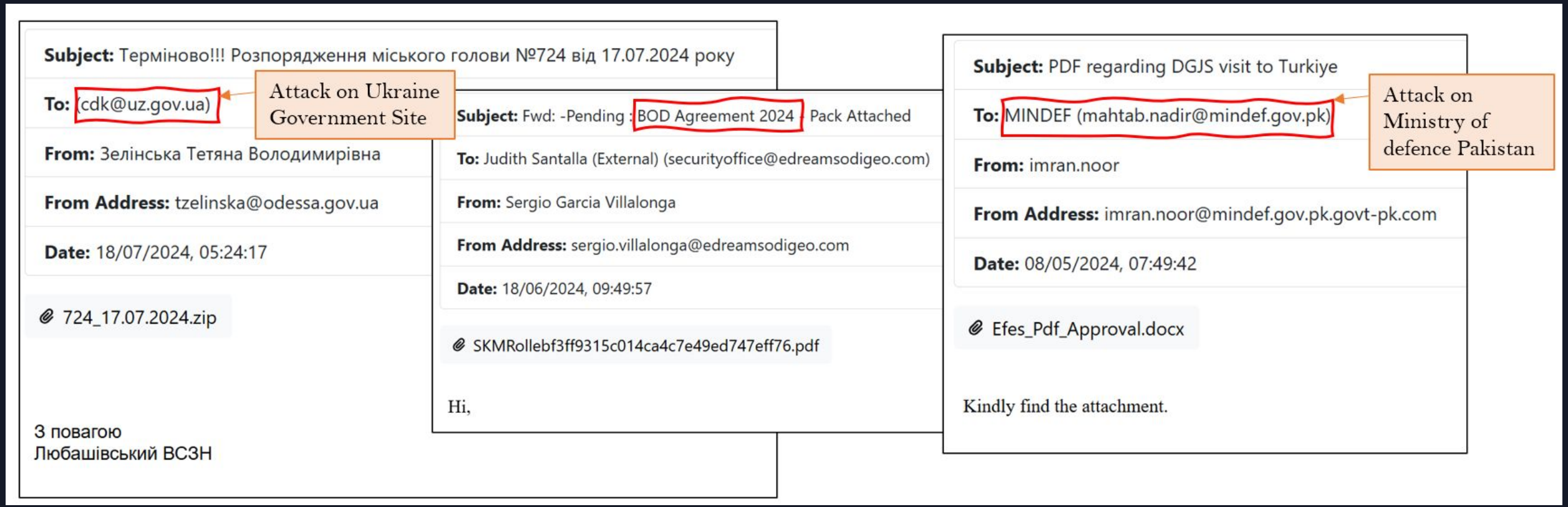
26% of phishing detected by NACE missed by ~95% of AV



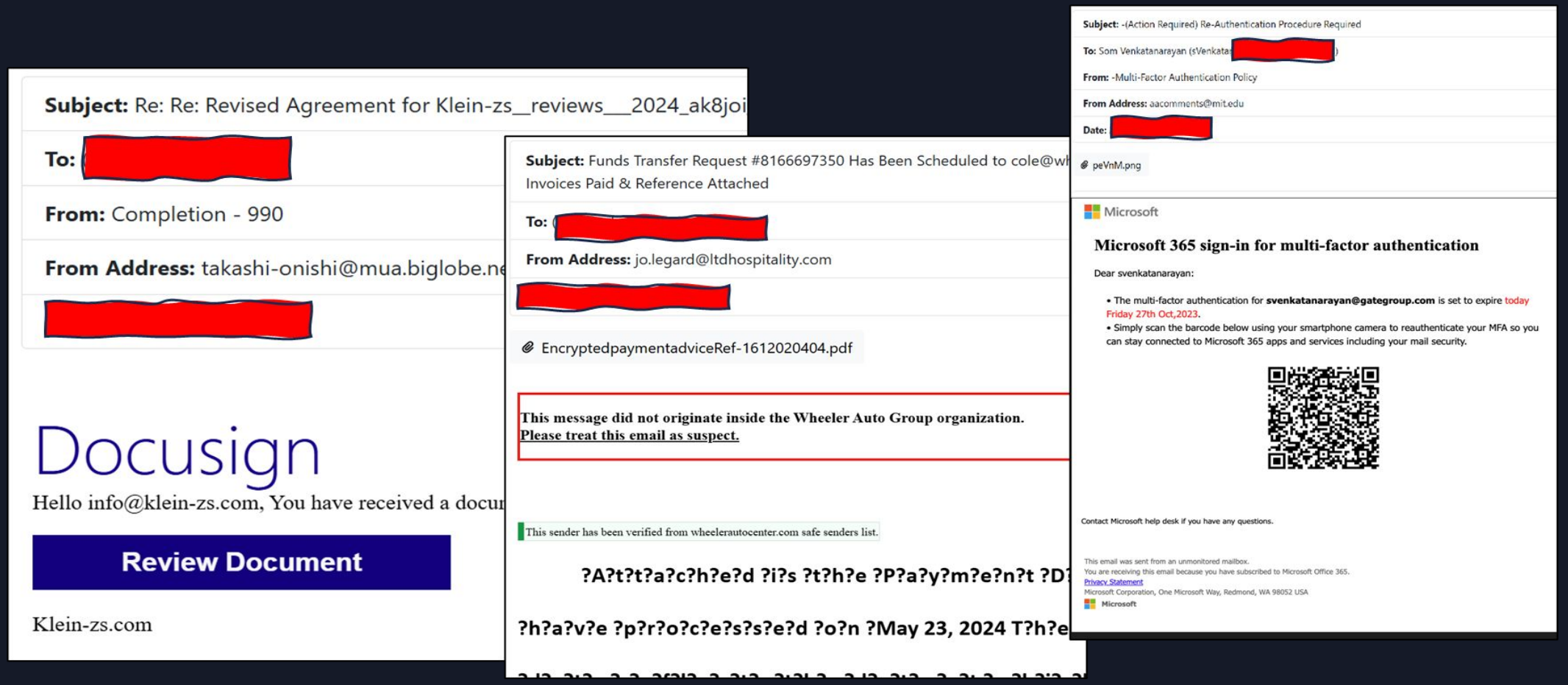
VirusTotal PDF Downloader Detection: Comparison of Other AV Vendors Against NACE's 100% Detection

54% of PDF downloaders detected by NACE missed by ~95%

NACE: Detection of Advanced Persistent Threats



NACE: Detection of Malicious Samples Unknown to AV



Summary

	Signature	Sandbox	ML applied on files	NGAV, EDR, XDR, MDR, Deception	Neural Analysis and Correlation Engine (NACE)
Evasive Threat Response	False Neg	Partial Detect / Respond	False Neg	Detect Remediate	Detect Prevent
	Multi-stage malware's first stage is benign, making detection challenging. Out of Scope - Identity based attacks - First stage of AI generated multi-stage malware having evasive malicious payloads.	Pre-filters optimizes scanning, allowing malicious attachments/ URLs to evade the sandbox. Never ending evasions to bypass sandbox (AI or non-AI generated), conceal the malicious payload leading to false negatives. Out of Scope - Identity based attacks.	Multi-stage malware's first stage is benign, making detection challenging. Out of Scope - Identity based attacks - First stage of AI generated multi-stage malware having evasive malicious payloads.	Post- execution detection. Dwell time, critical for response. Not every event can be extracted and sent to the cloud.	Learns from semantics and thematic structure embedded in emails to make decisions about attachments, independent of final malicious payload, landing URL for detection.
Identifies **Without** Malicious Payload/Landing URL?	No	No	No	No	Yes

ACKNOWLEDGEMENTS

The authors extend their gratitude to Andrey Yakymovych for his invaluable assistance in developing the framework for semantic extraction.