



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



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



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



The Hacker News

<https://thehackernews.com> › Cybersecurity News

New HTML Smuggling Campaign Delivers DCRat Malware ...

Sep 27, 2024 — Russian-speaking users have been targeted as part of a new campaign distributing a commodity trojan called DCRat (aka DarkCrystal RAT) by means of a technique ...



Recorded Future

<https://www.recordedfuture.com> › research › qr-code-a...

Security Challenges Rise as QR Code and AI-Generated ...

Jul 18, 2024 — QR code phishing, also known as "quishing," involves using manipulated or fake QR codes for malicious purposes. This technique has become ...



J.P. Morgan Private Bank

<https://privatebank.jpmorgan.com> › ... › Wealth Planning

Ransomware Attacks are increasingly sophisticated. Are ...

The rise and cost of a cyber ransom **In 2023, ransomware attacks impacted 1 in every 10** organizations worldwide, surging 33% from previous year.



Infosecurity Magazine

<https://www.infosecurity-magazine.com> › news › 341-rl...

Report Reveals 341% Rise in Advanced Phishing Attacks

May 22, 2024 — Security experts have reported a 341% increase in malicious phishing links, business email compromise (BEC), QR code and attachment-based threats in the past ...




The Hacker News

<https://thehackernews.com> › Cybersecurity News

PEAKLIGHT Downloader Deployed in Attacks Targeting ...

Aug 23, 2024 — New PEAKLIGHT PowerShell dropper, uncovered by Mandiant, deploys malware via fake movie downloads on Windows.

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 (...

<https://www.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

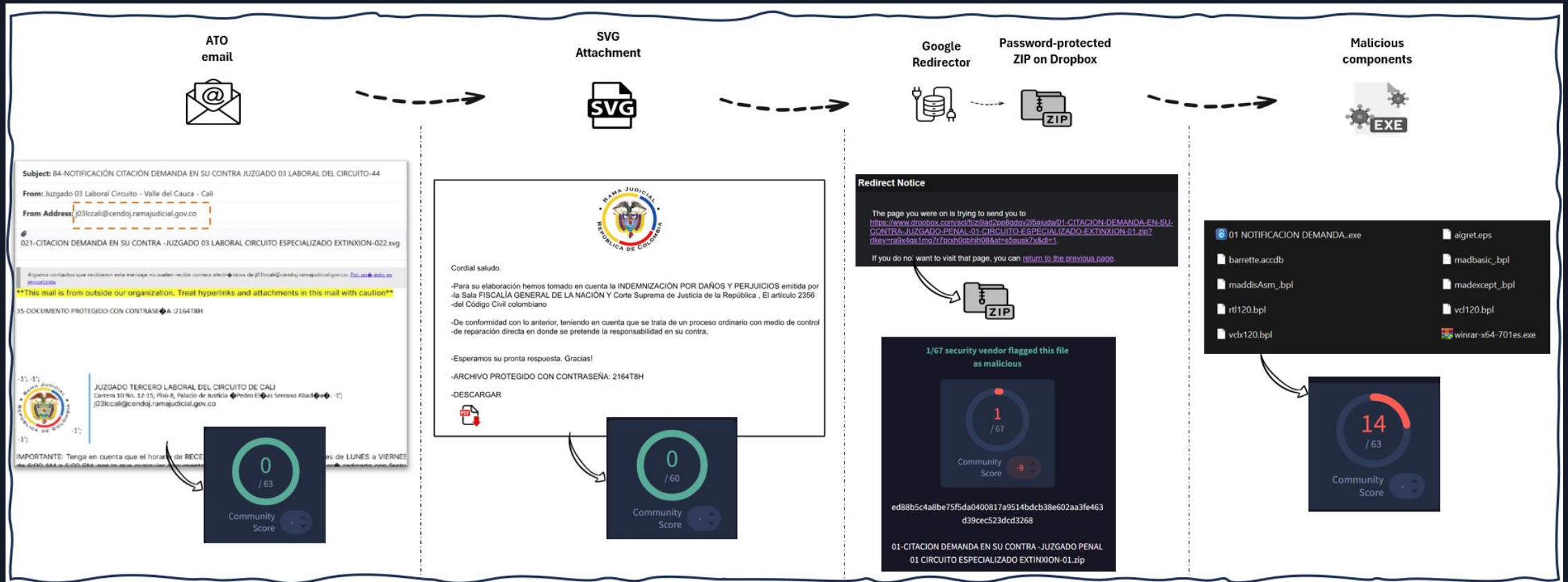
WhatsApp message: Caller ID +1 (408) 228-228, Arheart Your caller left you a message, Voice_459.html

HTML Attachment → Cloud Flare URL → Phishing URL

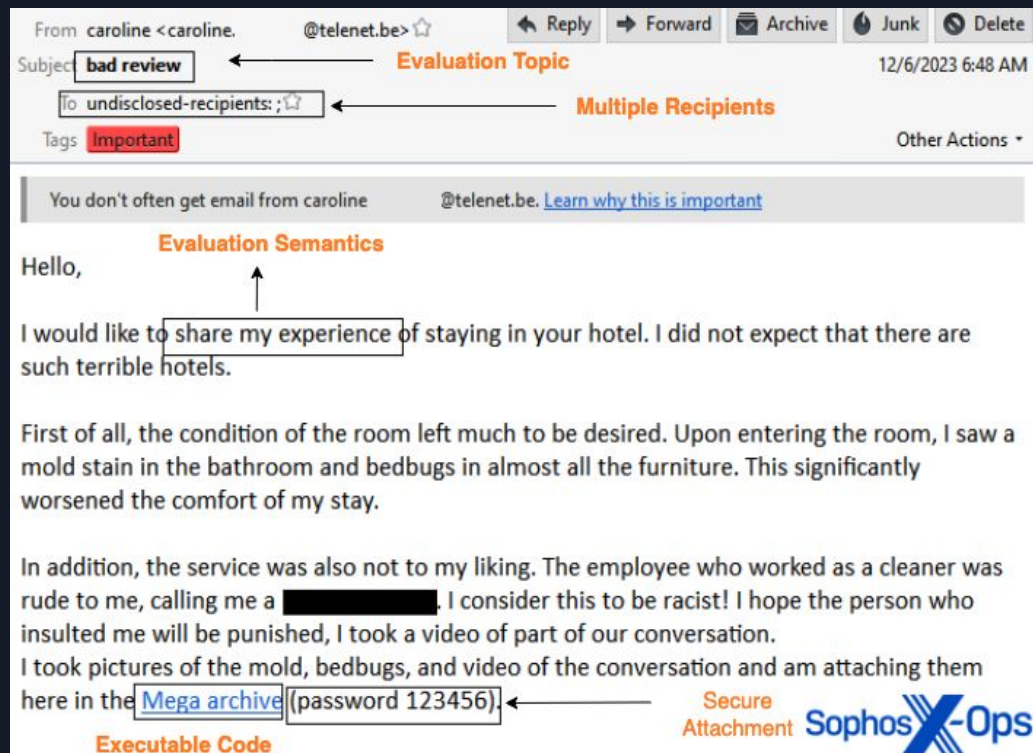
VirusShare analysis results:
1. 753d42902c846ee3b8af5c0ca2a8d41c76acd68f9bc6b8bbefe3ba7bd5c35065: 0 / 64, No security vendors and no sandboxes flagged this file as malicious.
2. 2613ed1a01711b3f67cc00c22f3a85775ae5fff036e079c1358882c9f5728ad4: 0 / 61, No security vendors and no sandboxes flagged this file as malicious.
3. 2613ed1a01711b3f67cc00c22f3a85775ae5fff036e079c1358882c9f5728ad4: 1 / 93, 1,93 security vendor flagged this URL as malicious. URL: https://hpjfq.com/?kgubvsqi, hpjfq.com, multiple-redirects.

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)

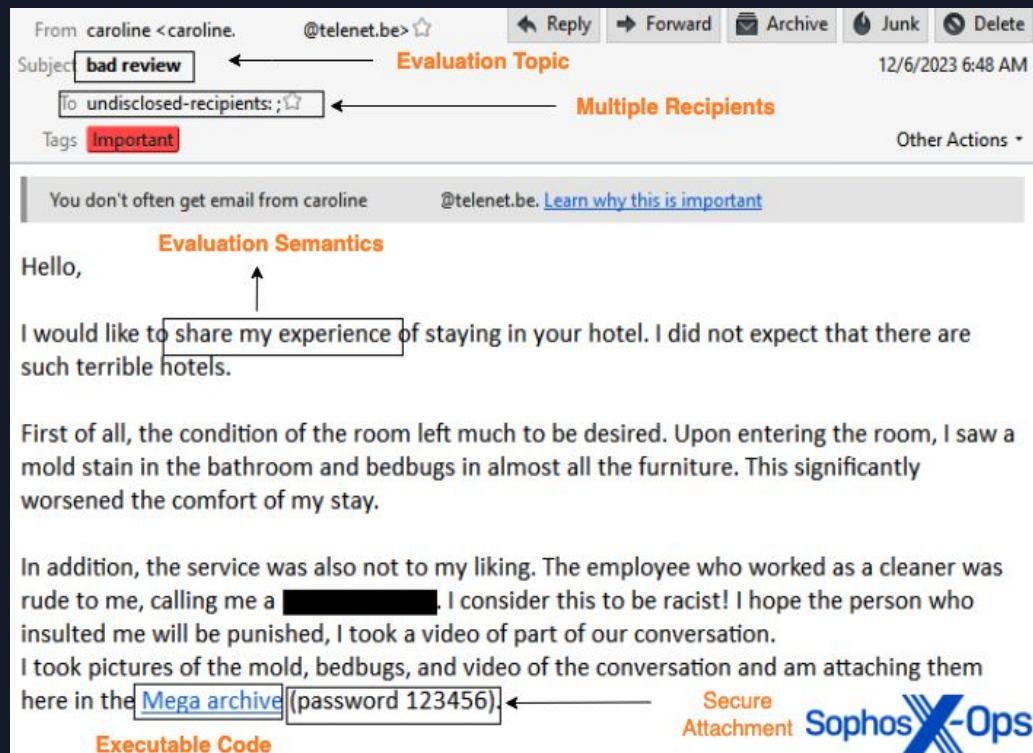
Therefore, evasions (human or AI) **Hide Malicious Payload** ⇒ Bypass Technology ⇒ Breach

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

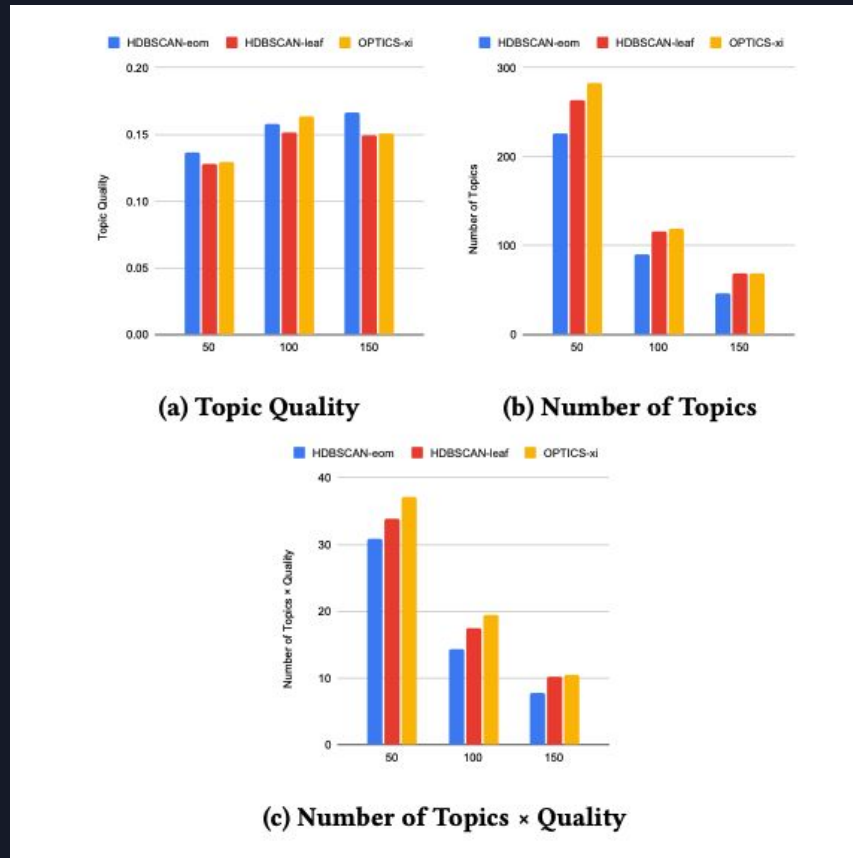
Design Steps : Semantics and Thematic Analysis for Classification

Step 1: Analyze Historic Threat Actor Emails

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

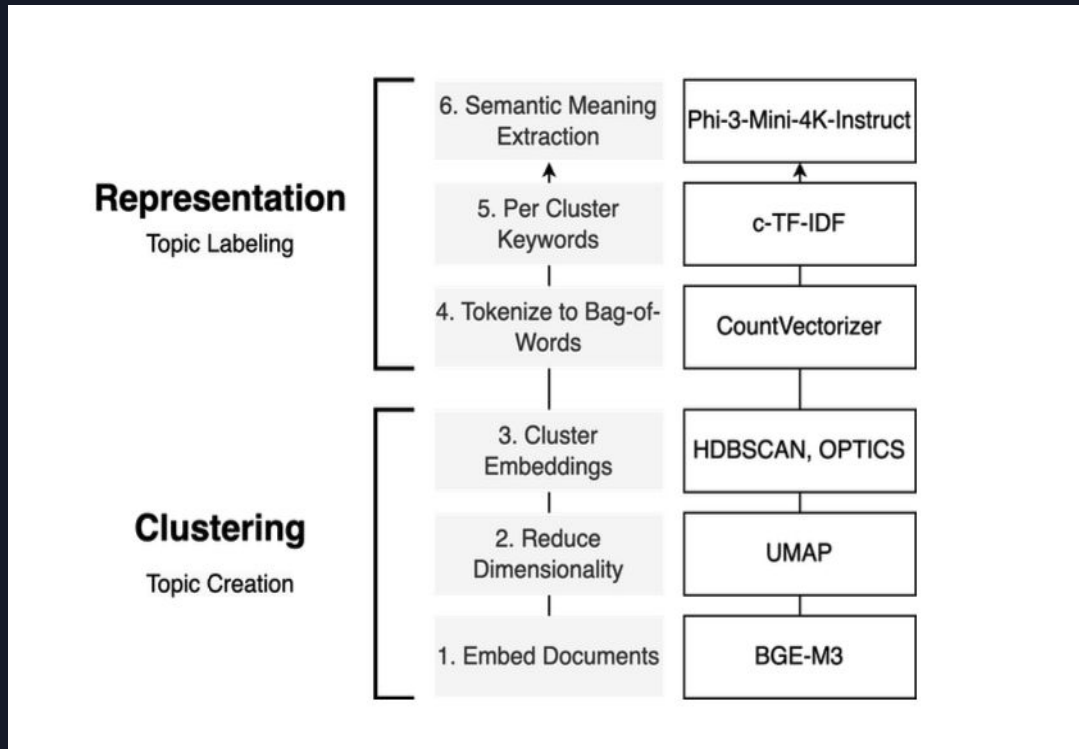
Becoming Immune to Evasion: Solving From First Principles

Experimentation with Unsupervised Clustering Algorithms



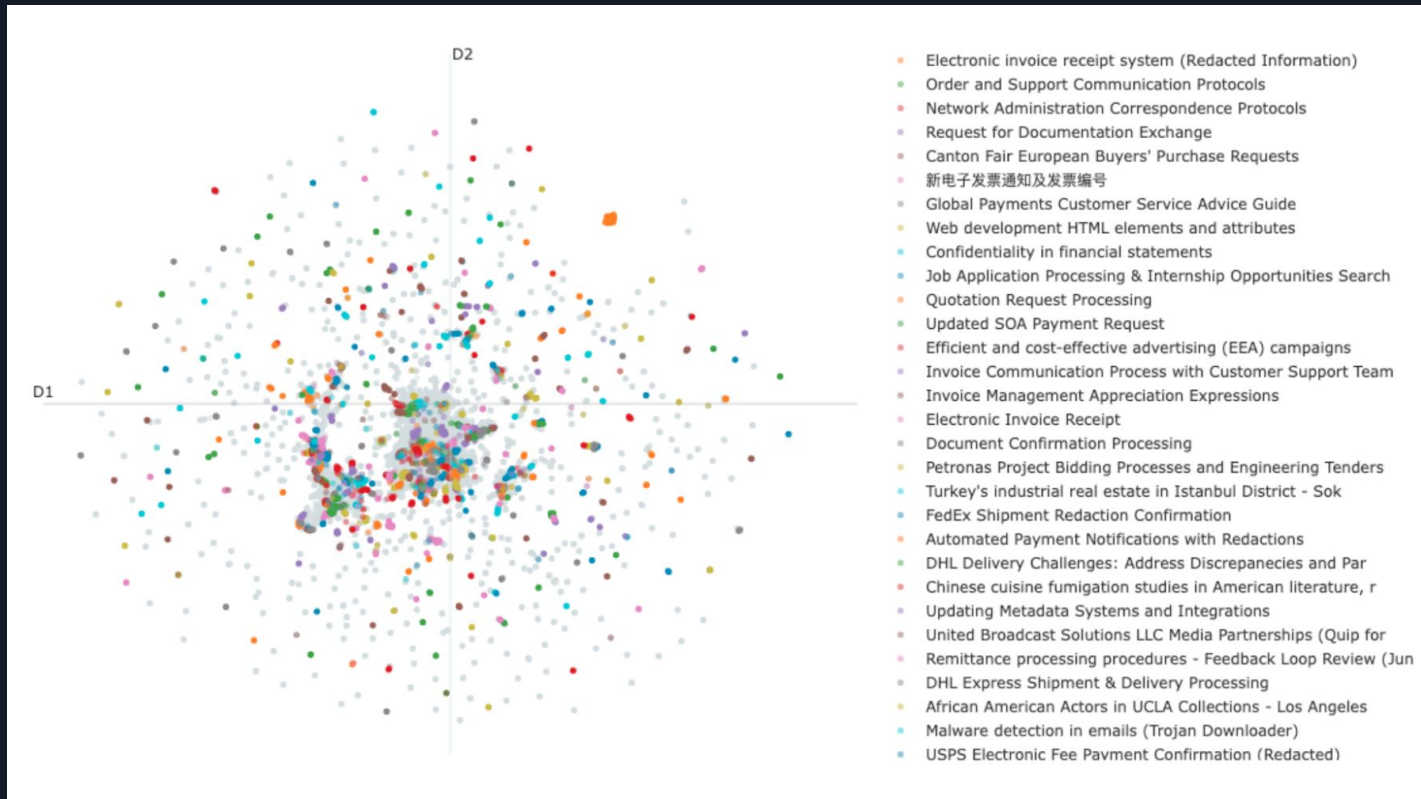
- Data Source:
 - Extracted email bodies, subject from historic emails used to deliver malware
- Algorithms Evaluated:
 - HDBSCAN - EOM
 - HDBSCAN - Leaf
 - OPTICS
- Evaluation Metrics:
 - Topic Quality - Measurement of granularity of cluster
 - Topic Coherence - Interpretability of a topic, closeness 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 250+ semantics**
Semantics which are extensively used by threat actors to deliver malicious attachments, 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.

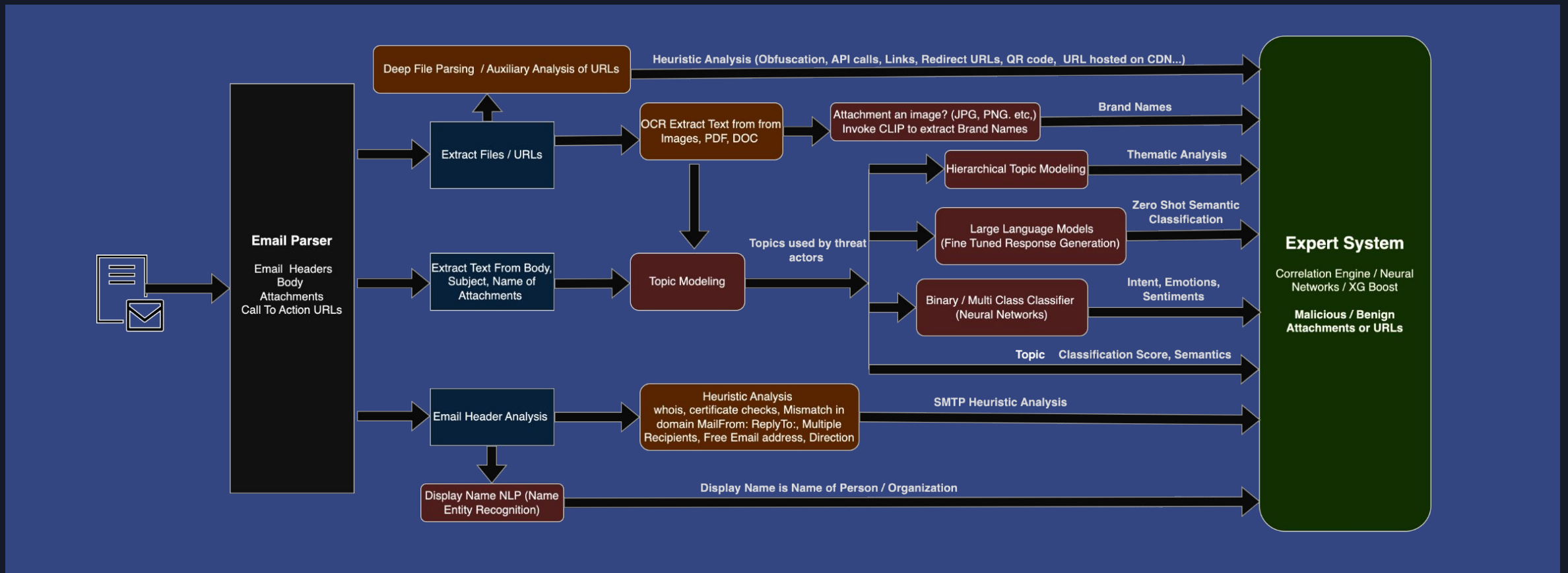
Step 2: Design an analysis system which does not need a malicious payload

- 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)

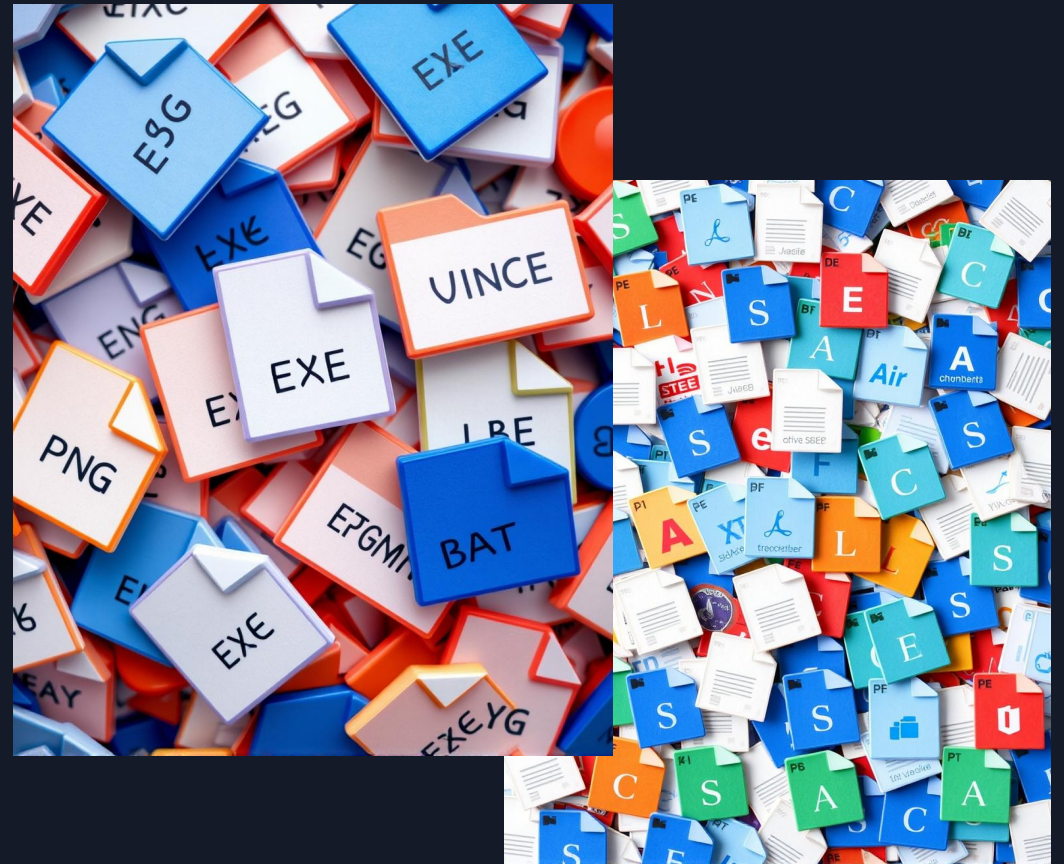
Leveraging Topics and Semantics to detect Malicious Attachments and URLs



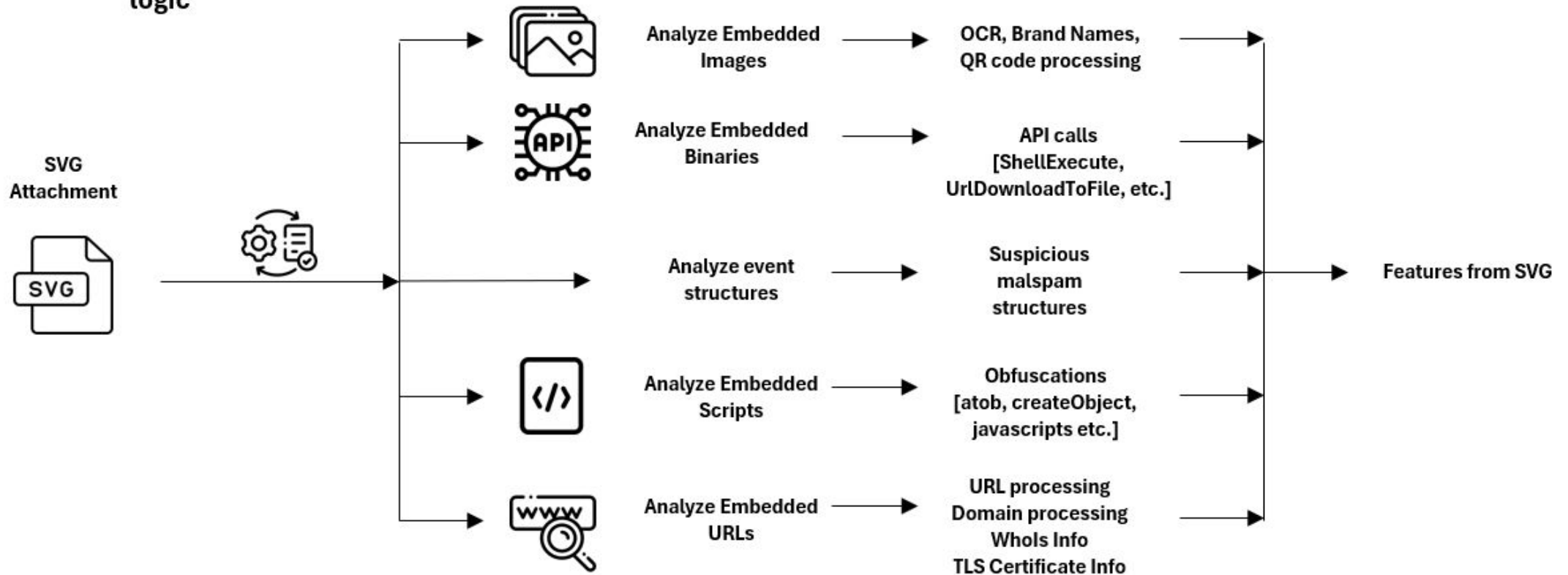
Deep File Analysis of NACE

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.)



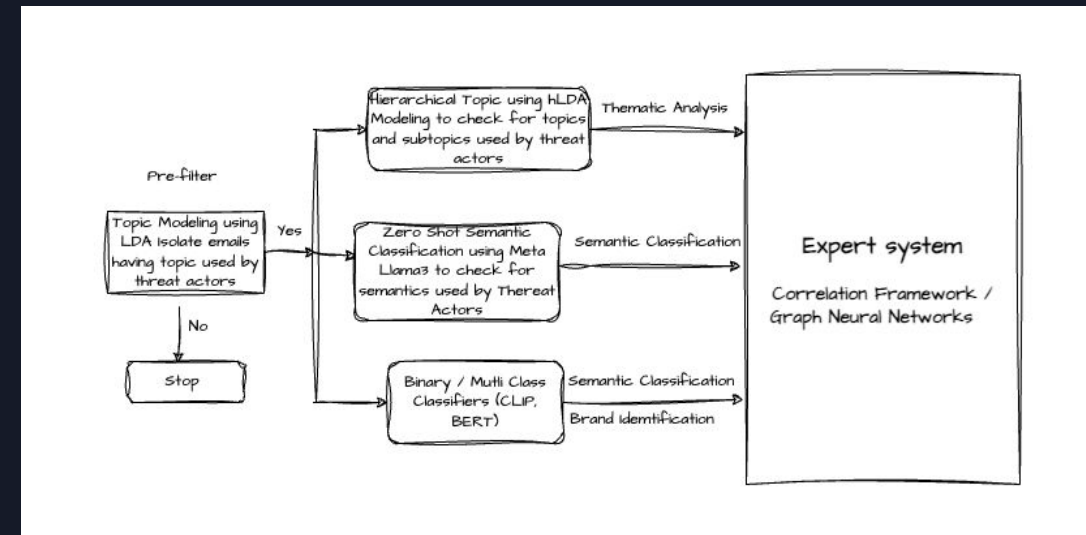
SVG: Deep file-parsing logic



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

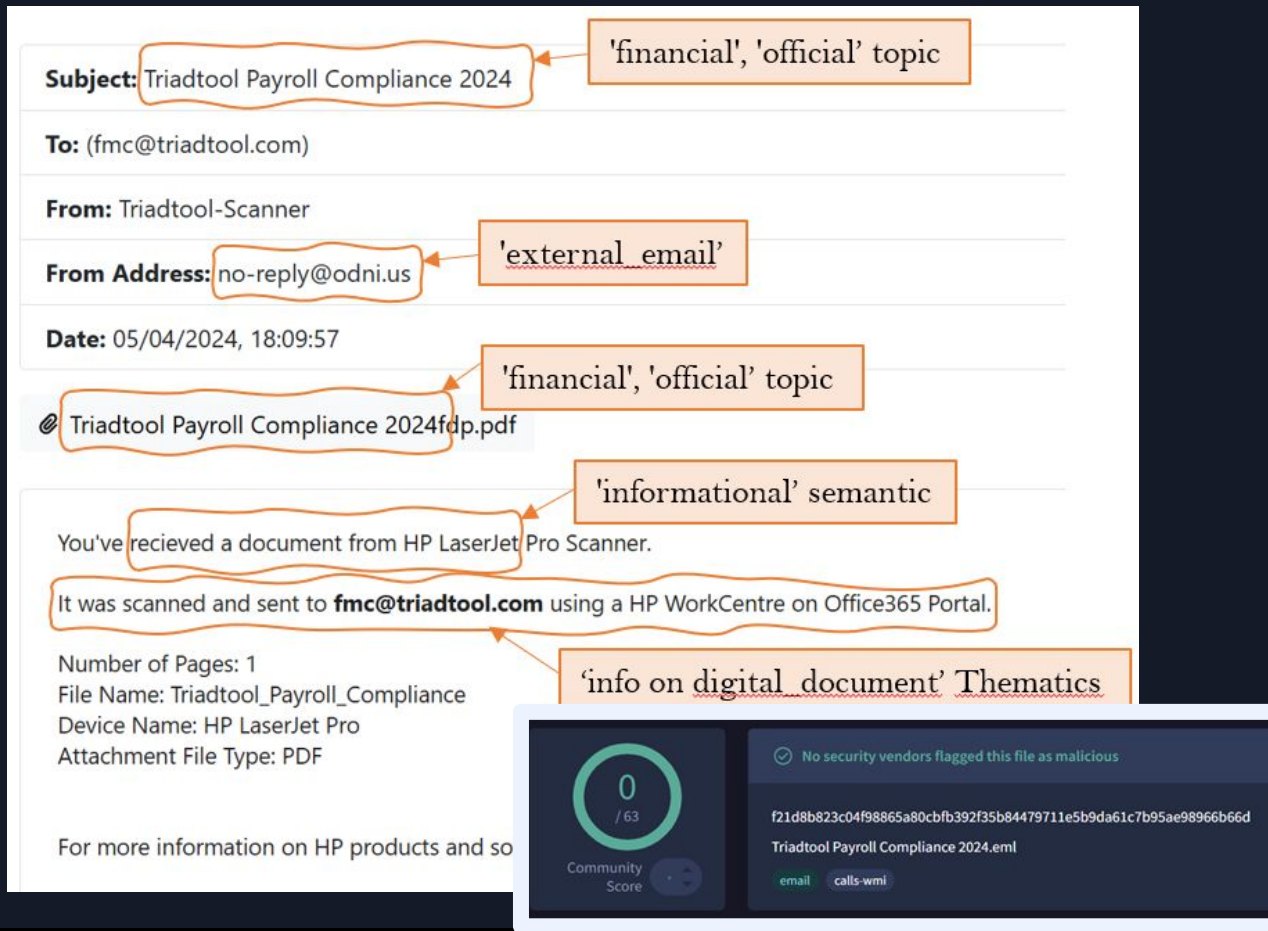
- Correlates Semantic analysis, Thematic Analysis, Topic Modeling, Deep File Parsing, SMTP Headers to decide if Attachment is malicious

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



Subject: Triadtool Payroll Compliance 2024 → 'financial', 'official' topic

To: (fmc@triadtool.com)

From: Triadtool-Scanner

From Address: no-reply@odni.us → 'external_email'

Date: 05/04/2024, 18:09:57

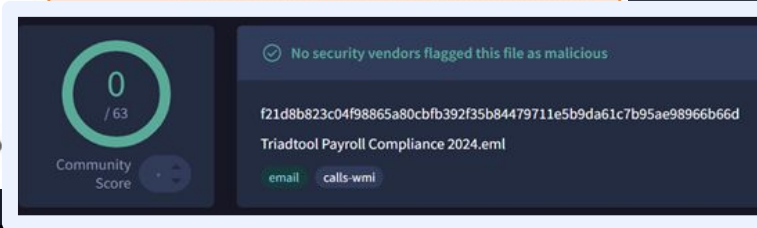
📎 Triadtool Payroll Compliance 2024fdp.pdf → 'financial', 'official' topic

You've received a document from HP LaserJet Pro Scanner. → 'informational' semantic

It was scanned and sent to **fmc@triadtool.com** using a HP WorkCentre on Office365 Portal. → 'info on digital document' Thematics

Number of Pages: 1
File Name: Triadtool_Payroll_Compliance
Device Name: HP LaserJet Pro
Attachment File Type: PDF

For more information on HP products and so



No security vendors flagged this file as malicious

f21d8b823c04f98865a80cbfb392f35b84479711e5b9da61c7b95ae98966b66d

Triadtool Payroll Compliance 2024.eml

email calls-wmi

Header Semantics

Sender Domain analysis:

- Whols 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



The screenshot shows an email from Triadtool with several analysis annotations:

- Check for known brand:** Points to the Triadtool logo.
- PDF_has_image_qr, PDF_has_url, PDF_has_single_page:** Points to the email content.
- 'financial', 'official' topic:** Points to the subject line "Payroll Compliance Request from Triadtool Human Resources".
- url_has_a_email, domain_has_numbers, domain_uses_cdn_service, domain_has_privacy_protection, cert_ca_letsencrypt:** Points to a QR code.
- URL:** <https://pub-52c3d43a0cfd46c3bd200403fcc586b4.r2.dev/payroll/0012qus339.html#fmc@triadtool.com>
- Security Scan:** A security vendor scan shows a score of 0/61 and identifies the file as "Triadtool Payroll Compliance 2024fdp.pdf".

Attachment Semantics

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

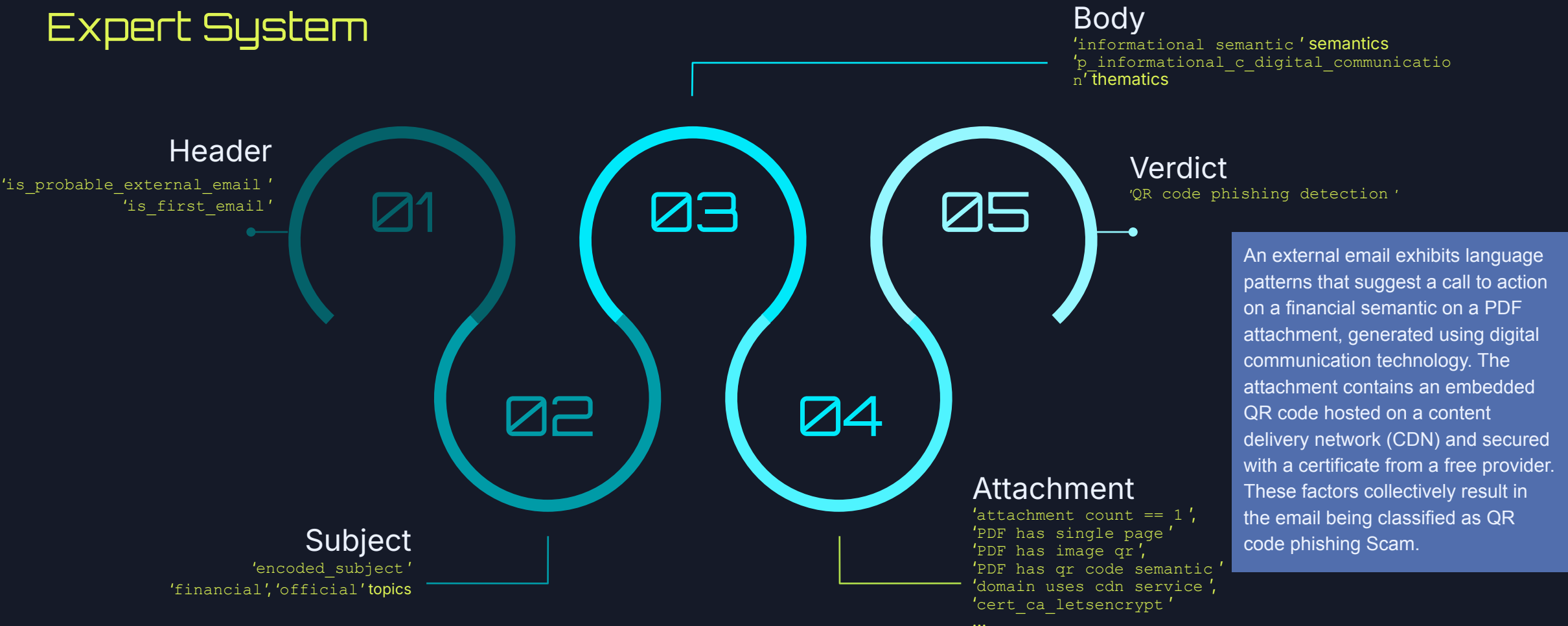
Image Processing

- OCR, Brand Identification, QR Code Identification, etc.

URL Analysis

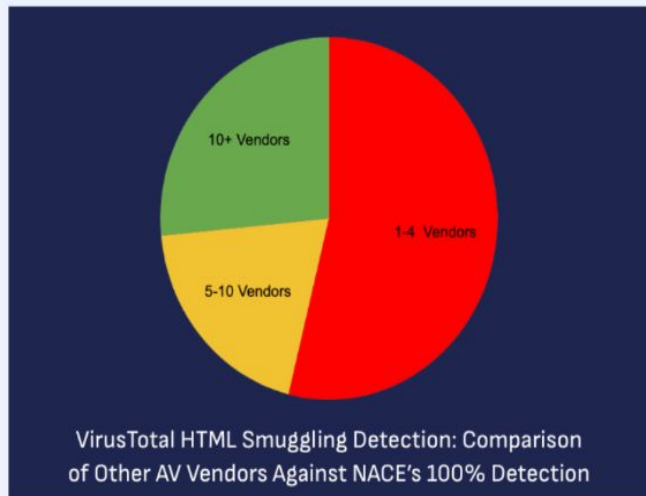
- Suspicious URL format analysis
- Identification of suspicious domains
- Whois information
- Domain Creation info
- Domain Categories
- TLS Certificates analysis
- Analyzing Document sent as URLs
- ...

Case-Study Continued: Expert System

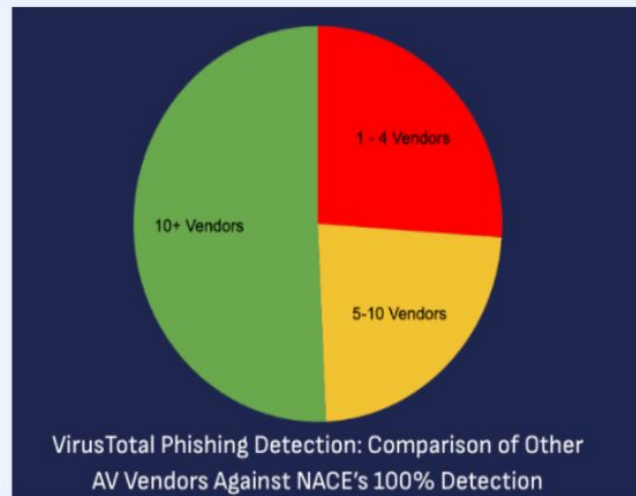


Benchmark against other Technologies

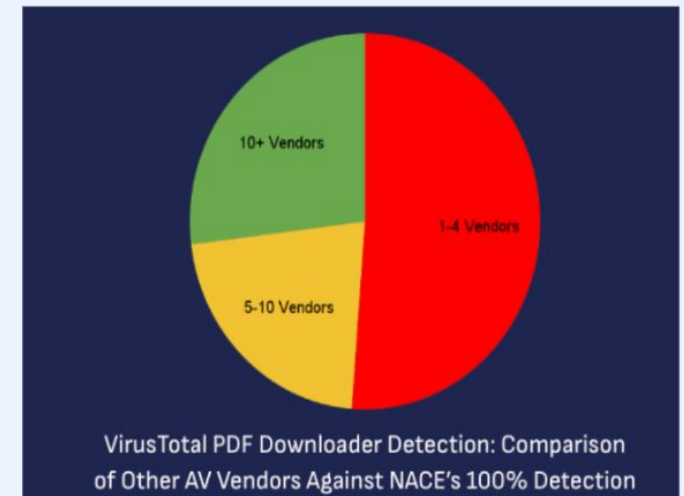
- **Data Source:** ~13K Samples, 2024 Evasive threats (HTML Smuggling, Phishing, Downloaders, Dropper, etc) from Viru Total
- **VirusTotal:** 70 AV Vendors, 70 URL Black List, 10 Sandboxes
- **Results:** 99% of coverage, ~44% of the evasive threats detected by NACE were missed by 95% of the AV Technologies



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



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



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

NACE: Detection of Advanced Persistent Threats

Subject: Терміново!!! Розпорядження міського голови №724 від 17.07.2024 року

To: (cdk@uz.gov.ua)

Attack on Ukraine
Government Site

From: Зелінська Тетяна Володимирівна

From Address: tzelinska@odessa.gov.ua

Date: 18/07/2024, 05:24:17

📎 724_17.07.2024.zip

З повагою
Любашівський ВСЗН

Subject: Fwd: -Pending : BOD Agreement 2024 Pack Attached

To: Judith Santalla (External) (securityoffice@edreamsodigeo.com)

From: Sergio Garcia Villalonga

From Address: sergio.villalonga@edreamsodigeo.com

Date: 18/06/2024, 09:49:57

📎 SKMRollebf3ff9315c014ca4c7e49ed747eff76.pdf

Hi,

Subject: PDF regarding DGJS visit to Turkiye

To: MINDEF (mahtab.nadir@mindef.gov.pk)

Attack on
Ministry of
defence Pakistan

From: imran.noor

From Address: imran.noor@mindef.gov.pk.govt-pk.com

Date: 08/05/2024, 07:49:42

📎 Efes_Pdf_Approval.docx

Kindly find the attachment.

NACE: Detection of Malicious Samples Unknown to AV

Subject: Re: Re: Revised Agreement for Klein-zs__reviews__2024_ak8joi

To: [REDACTED]

From: Completion - 990

From Address: takashi-onishi@mua.biglobe.ne.jp

[REDACTED]

DocuSign
Hello info@klein-zs.com, You have received a document.

[Review Document](#)

Klein-zs.com

Subject: Funds Transfer Request #8166697350 Has Been Scheduled to cole@w...
Invoices Paid & Reference Attached

To: [REDACTED]

From Address: jo.legard@ltdhospitality.com

[REDACTED]

📎 EncryptedpaymentadviceRef-1612020404.pdf

**This message did not originate inside the Wheeler Auto Group organization.
Please treat this email as suspect.**

This sender has been verified from wheelerautocenter.com safe senders list.

?A?t?t?a?c?h?e?d ?i?s ?t?h?e ?P?a?y?m?e?n?t ?D?

?h?a?v?e ?p?r?o?c?e?s?s?e?d ?o?n ?May 23, 2024 T?h?e

Subject: -(Action Required) Re-Authentication Procedure Required

To: Som Venkatanarayan (sVenkata [REDACTED])

From: -Multi-Factor Authentication Policy

From Address: aacomment@mit.edu

Date: [REDACTED]

📎 peVnM.png

Microsoft

Microsoft 365 sign-in for multi-factor authentication

Dear svenkatanarayan:

- The multi-factor authentication for **svenkatanarayan@gategroup.com** is set to expire **today Friday 27th Oct, 2023**.
- Simply scan the barcode below using your smartphone camera to reauthenticate your MFA so you can stay connected to Microsoft 365 apps and services including your mail security.



Contact Microsoft help desk if you have any questions.

This email was sent from an unmonitored mailbox.
You are receiving this email because you have subscribed to Microsoft Office 365.
[Privacy Statement](#)
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA
Microsoft

Summary

	Signature	Sandbox	ML applied on files	NGAV, EDR, XDR, MDR, Deception	Neural Analysis and Correlation Engine (NACE)
Evasive Threat Response	<p>False Neg</p> <p>Multi-stage malware's first stage is benign, making detection challenging.</p> <p>Out of Scope</p> <ul style="list-style-type: none"> - Identity based attacks - First stage of AI generated multi-stage malware having evasive malicious payloads. 	<p>Partial</p> <p>Detect Respond</p> <p>Pre-filters optimizes scanning, allowing malicious attachments/ URLs to evade the sandbox.</p> <p>Never ending evasions to bypass sandbox (AI or non-AI generated), conceal the malicious payload leading to false negatives.</p> <p>Out of Scope</p> <ul style="list-style-type: none"> - Identity based attacks. 	<p>False Neg</p> <p>Multi-stage malware's first stage is benign, making detection challenging.</p> <p>Out of Scope</p> <ul style="list-style-type: none"> - Identity based attacks - First stage of AI generated multi-stage malware having evasive malicious payloads. 	<p>Detect</p> <p>Remediate</p> <p>Post- execution detection.</p> <p>Dwell time, critical for response.</p> <p>Not every event can be extracted and sent to the cloud.</p>	<p>Detect</p> <p>Prevent</p> <p>Learns from semantics and thematic structure embedded in emails to make decisions about attachments, independent of final malicious payload, landing URL for detection.</p>
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.