ThreatIntelGPT: An AI-Powered Cyber Threat Intelligence System Using NLP, IOC Extraction, MITRE ATT&CK Mapping and CVE Explanation

Mala Neeraj Srinivas
Department of Computer Science and Engineering
Indian Institute of Information Technology, Surat
Surat, India
Email: ui21cs35@iiitsurat.ac.in

***Abstract*—The accelerating frequency of cyber attacks has led to increasing volumes of Cyber Threat Intelligence (CTI) data published by security agencies and researchers. These reports contain unstructured narratives describing attack vectors, Indicators of Compromise (IOCs) and vulnerabilities, making manual extraction slow and expertise-dependent. This paper presents ThreatIntelGPT, an AI-driven CTI processing system that automates the ingestion and summarisation of threat feeds using Natural Language Processing (NLP). The system performs RSS ingestion, IOC extraction, Named Entity Recognition (NER), MITRE ATT&CK mapping and AI-powered CVE explanation. A lightweight voice assistant interface enables hands-free querying of CTI information. ThreatIntelGPT is implemented using FastAPI, spaCy, a rule-based IOC engine, NVD CVE lookup and SQLite storage. Evaluation shows that the system reduces analyst workload, improves triage efficiency and provides a scalable solution for SOC environments.**

***Index Terms*—Cyber Threat Intelligence, NLP, IOCs, MITRE ATT&CK, CVE Analysis, FastAPI, Automation.**

I. Introduction

Cybersecurity organisations must analyse threat intelligence reports to detect malicious activity and understand emerging threats. Authoritative sources such as CISA, private researchers and security vendors publish numerous CTI articles daily. These reports contain essential information such as IOCs, exploited vulnerabilities and threat actor tactics.

However, CTI is largely unstructured, making manual analysis slow and error-prone. Moreover, SOC teams are often overwhelmed by large volumes of incoming alerts and lack the bandwidth to manually inspect every intelligence report. Automated CTI processing enables faster decision-making and reduces analyst fatigue, especially in environments where rapid triage is crucial.

To address this challenge, we introduce **ThreatIntelGPT**, an automated NLP-driven pipeline that processes CTI feeds and extracts actionable insights. Key contributions include automated ingestion and summarisation, regex-based IOC extraction, spaCy NER, MITRE ATT&CK mapping, CVE analysis with AI explanation and a browser-based voice assistant. The system is designed for low-resource environments and can run on standard CPU hardware without requiring GPU acceleration.

II. System Architecture

ThreatIntelGPT is a modular architecture consisting of four major layers: the Data Ingestion Layer, Processing Layer, Storage Layer and User Interaction Layer. The separation ensures scalability and easy debugging, and allows additional NLP modules to be integrated later.

A. 1) Data Ingestion Layer

This layer collects CTI content from RSS feeds such as CISA, KrebsOnSecurity, The Hacker News and NVD CVE feeds. For short summaries, additional article text is scraped. This helps improve the quality of analysis, as many RSS feeds truncate their descriptions.

B. 2) Processing Layer

The FastAPI backend performs multiple NLP tasks in sequence:

- text normalisation,
- lightweight summarisation,
- IOC extraction,
- NER via spaCy,
- MITRE ATT&CK mapping,
- CVE enrichment with AI-generated explanation.

Each module is isolated into separate Python files, ensuring modularity and enabling researchers to swap components (e.g., upgrading the summariser or NER model later).

C. 3) Storage Layer

SQLite stores article titles, summaries, IOCs, entities, MITRE mappings and raw CTI text. Storing both the summary and full text enables future reprocessing if more advanced NLP modules are added.

D. 4) Interaction Layer

The frontend dashboard (HTML/CSS/JS) displays CTI summaries, MITRE mappings and CVE insights. It shows statistics, saved reports, IOC lists and extracted entities. A Level-1 voice assistant uses the Web Speech API for browser-based speech capture, enabling future integration of real speech-to-text pipelines.

Additionally, the UI is optimised to operate as a local SOC dashboard, emphasising readability, dark-mode design and real-time interaction.

III. NLP Pipeline Components

A. RSS Ingestion and Preprocessing

Text is normalised by lowercasing, HTML stripping, cleaning non-UTF symbols and expanding truncated summaries. The system also removes boilerplate content and irrelevant metadata often present in CTI feeds.

B. IOC Extraction

A rule-based regex engine identifies IPv4/IPv6 addresses, domain names, URLs and cryptographic hashes. It was manually tuned based on common CTI patterns, and avoids overmatching by applying strict boundary conditions.

C. Named Entity Recognition

spaCy's `en_core_web_sm` model extracts organisations, locations, malware names and threat actor entities. Although this model is lightweight, it provides sufficiently accurate results for SOC-level CTI triage. The modular design allows the use of larger transformer-based NER models in the future.

D. MITRE ATT&CK Mapping

Keywords such as "persistence", "command execution" and "credential dumping" are mapped to ATT&CK techniques including T1059, T1105 and T1003. The mapping dictionary is expandable, allowing continuous updates as adversary behaviours evolve.

E. CVE Lookup and Explanation

The system queries NVD's public API, extracts severity and CVSS scores, and generates AI-based explanations using HuggingFace models with fallback logic. The fallback generator ensures that users always receive a useful summary even when the AI model is offline or API limits are reached.

F. Limitations

Transformer-based summarisation is not used due to CPU constraints, and the Web Speech API currently supports only client-side speech capture. However, both of these limitations are addressed in the future work section.

TABLE I
Evaluation Results on 21 CTI Articles

Metric	Result
Average ingestion time	1.8 sec/article
IOC extraction precision	92%
NER accuracy	87%
MITRE mapping accuracy	78%
CVE explanation reliability	100% (fallback supported)

IV. Evaluation

ThreatIntelGPT was tested on 21 CTI articles. Table I summarises key results:

Analysts noted reduced reading time (60–70%), faster IOC extraction and improved threat classification. The system also demonstrated reliable performance on CPU-only machines, confirming its suitability for low-resource deployments.

Furthermore, the voice assistant was positively reviewed for accessibility, especially in situations where analysts may be multitasking or inspecting logs while needing quick CTI summaries.

V. Conclusion

ThreatIntelGPT provides an efficient automated CTI pipeline combining NLP, MITRE mapping and CVE analysis into a unified system. Its lightweight design enables deployment on standard CPU hardware, benefiting security teams and researchers. The system bridges the gap between raw CTI feeds and actionable intelligence, enabling faster incident response and reducing manual analysis overhead.

VI. Future Work

Planned improvements include:

- Full speech-to-text backend for real conversational interaction,
- Transformer-based summarisation (e.g., T5, BART),
- Threat clustering and graph-based correlation,
- SIEM integration with Splunk or Elastic,
- Real-time alerting and incident response automation.

References

[1] ExplosionAI, "spaCy Industrial NLP Library," 2024.
[2] MITRE Corporation, "MITRE ATT&CK Framework," 2024.
[3] HuggingFace, "Transformers and Inference API," 2024.
[4] CISA, "Cybersecurity Advisories," https://cisa.gov.
[5] FastAPI, "FastAPI Framework," https://fastapi.tiangolo.com.