From Data Overload to Clear Insights: Tackling Vulnerabilities with AI

1st Vaishnavi Sanjay Sadul
College of Engineering(MSDAE)
Northeastern University
Vancouver, Canada
sadul.v@northeastern.edu

2nd Dhaval Jariwala
College of Engineering(MSDAE)
Northeastern University
Vancouver, Canada
jariwala.dh@northeastern.edu

3rdYihua Cai
College of Engineering(MSDAE)
Northeastern University
Vancouver, Canada
cai.yihu@northeastern.edu

4th Yuze Li
College of Engineering (MSDAE)
Northeastern University
Vancouver, Canada
li.yuze3@northeastern.edu

***Abstract*—Understanding cybersecurity data often feels overwhelming due to the sheer complexity of the information. This paper presents a streamlined solution leveraging artificial intelligence (AI) to transform data from the National Vulnerability Database (NVD) into actionable insights. Key contributions include a custom Named Entity Recognition (NER) pipeline tailored to cybersecurity datasets, an interactive dashboard for visualizing trends, and a knowledge graph for revealing hidden relationships among vulnerabilities. Our approach enhances the ability to identify and manage risks efficiently, significantly outperforming generic AI models in identifying critical patterns.**

***Index Terms*—Cybersecurity Data, National Vulnerability Database (NVD), Named Entity Recognition (NER), Vulnerability Analysis, Knowledge Graph, Risk Management**

I. Introduction

A. Context and Motivation

The rapid growth of digital infrastructure has brought about an exponential increase in cybersecurity threats, making it imperative for organizations to identify and address vulnerabilities effectively. The National Vulnerability Database (NVD) is a vital resource, providing extensive information on vulnerabilities such as the Known Exploited Vulnerabilities (KEV) and Common Vulnerabilities and Exposures (CVE). However, the unstructured and voluminous nature of this data often overwhelms analysts, hindering timely decision-making. Traditional approaches to exploring this data involve manual processes or reliance on generic tools that fail to address the domain-specific intricacies of cybersecurity.

This project seeks to bridge the gap between raw data and actionable insights by developing an AI-driven system. By integrating domain-specific Named Entity Recognition (NER) models, interactive dashboards, and knowledge graphs, we aim to enable more efficient and intuitive exploration of cybersecurity vulnerabilities. The proposed system is designed to tackle key challenges such as the extraction of critical information, identification of hidden relationships, and visualization of complex patterns within the data. This work aligns with the growing need for solutions that transform overwhelming datasets into meaningful, structured outputs that drive informed decision-making.

By focusing on domain-specific AI solutions, the project seeks to address limitations of existing generic tools and enhance the accuracy and utility of cybersecurity data analysis. The NER models are tailored to recognize and extract cybersecurity-specific entities, such as vulnerability names and descriptions, with greater precision. The interactive dashboards simplify the visualization of trends and patterns, allowing stakeholders to quickly identify priority areas. Additionally, the knowledge graph enables users to explore connections and clusters, revealing risk pathways and dependencies that would be difficult to discern in traditional spreadsheet-based analyses.

In doing so, the project not only simplifies the workflow for analysts but also empowers decision-makers with actionable insights. By reducing the manual effort required and improving the accessibility of complex cybersecurity data, this system offers a practical and scalable solution for organizations to proactively manage vulnerabilities and mitigate risks. As cyber threats continue to evolve, the development of such tools is crucial in supporting a safer and more secure digital environment.

B. Research Objectives

- **Enhance Named Entity Recognition (NER):** Develop a custom NER pipeline optimized for cybersecurity datasets like KEV and CVE, improving precision, recall, and F1-score over generic models [1], [2].
- **Develop Interactive Dashboards:** Introduce user-friendly dashboards to visualize trends and patterns, enabling both technical and non-technical stakeholders to extract actionable insights [3], [4].

- **Implement Knowledge Graphs:** Design an interactive knowledge graph to reveal relationships among vulnerabilities, suppliers, and systems, facilitating proactive risk assessment [2].
- **Focus on Domain-Specific AI Models:** Demonstrate the value of tailored AI tools for addressing the unique challenges of cybersecurity data, building on prior work [5], [6].
- **Improve Decision-Making:** Provide a comprehensive framework that bridges the gap between raw data and informed risk management, enhancing accessibility and usability [3], [4].

C. Scope and Structure

The outline of the paper is as follows. We first discuss related works in AI-driven cybersecurity analysis and risk management in Section II. We define the challenges in analyzing and interpreting unstructured vulnerability data in Section III. In Section IV, we describe our methodology, including the development of a custom Named Entity Recognition (NER) pipeline, an interactive dashboard, and a knowledge graph. We evaluate the framework's performance in Section V and present key results. Finally, we summarize our findings and highlight future research directions in Section VI.

II. Literature Survey

A. Data-Driven and Knowledge-Based NER Approaches

Gao et al. (2021) proposed a hybrid Named Entity Recognition (NER) framework that integrates rule-based methods with machine learning to improve cybersecurity data extraction. The method employs:

1) **Domain-Specific Rules**: These rules enhance the recognition of rare and technical entities, such as vulnerability identifiers and attack vectors, which are commonly found in cybersecurity reports like CVE summaries.
2) **Pre-trained Word Embeddings**: By leveraging GloVe embeddings fine-tuned with cybersecurity-related data, the model achieves better semantic understanding of specialized terms.

While this method demonstrates strong performance on structured datasets, its reliance on predefined rules makes it rigid and less effective in dynamic contexts where new threats and patterns frequently emerge. Additionally, the lack of real-time adaptability limits its utility in scenarios requiring continuous monitoring. This research extends these concepts by integrating an iterative annotation process with custom NER models trained on KEV datasets to dynamically adapt to evolving threats.

B. Transformer-Based NER Models for Threat Intelligence

Alam et al. (2022) introduced **CyNER**, a Python-based library that utilizes transformer architectures such as BERT for extracting cybersecurity-specific entities, including Indicators of Compromise (IoCs). Its key innovations include:

1) **Contextualized Embeddings**: Transformer-based models capture contextual relationships between entities, enabling high accuracy in identifying malware names, IP addresses, and attack techniques from textual data.
2) **Heuristic Integration**: By combining heuristic filtering with machine learning, CyNER enhances its ability to detect patterns that are otherwise overlooked by purely data-driven approaches.

While CyNER performs well on static, structured datasets, it lacks adaptability for specific organizational needs, such as prioritizing high-risk vulnerabilities in large-scale, real-time environments. Moreover, the library does not integrate results into actionable dashboards, which limits its application for operational teams. This research addresses these gaps by tailoring NER models to specific domains and integrating the results into a real-time Streamlit-powered dashboard that allows users to explore risk types and connections interactively.

C. Enhancing NER Robustness with Advanced Learning Techniques

A 2023 study introduced **JCLB**, a dual-layer NER model combining contrastive learning and belief rule bases. This approach focuses on improving model robustness in handling noisy and ambiguous cybersecurity data. Key contributions include:

1) **Contrastive Learning**: By optimizing text embeddings, this technique enhances the model's ability to distinguish between similar entities, improving precision in identifying overlapping terms like malware names and exploit types.
2) **Belief Rule Bases**: Domain-specific rules are used to disambiguate entities, particularly in complex threat intelligence contexts where entity relationships are critical.

JCLB demonstrated excellent performance in structured datasets but struggles with unstructured and multimodal data, such as network logs and real-time alerts. Additionally, the model's reliance on batch processing limits its use in time-sensitive cybersecurity operations. This research builds upon JCLB by incorporating multimodal data sources, including KEV and CVE datasets, and adapting NER for real-time applications. The use of iterative testing and refinement ensures the model achieves higher F1-scores in dynamic scenarios.

D. Domain-Specific NER for Cyber Threat Intelligence Extraction

- CyTIE (2023) focuses on extracting actionable cybersecurity threat intelligence, such as malware groups, attack methods, and vulnerability information, from textual data. The model employs:
  1) **Custom NER Pipelines**: These pipelines are designed specifically for cybersecurity domains, ensuring accurate identification of key entities and their interrelationships.
  2) **Pattern Recognition**: By analyzing co-occurrence patterns, CyTIE uncovers hidden connections between attack techniques and vulnerabilities, enabling better prediction and response.

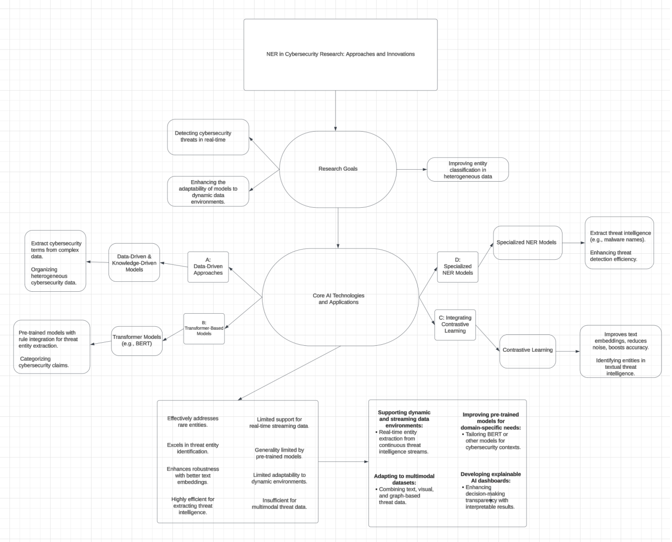

NER in Cybersecurity Research: Approaches and Innovations
Research Goals
Detecting cybersecurity threats in real-time
Enhancing the adaptability of models to dynamic data environments.
Improving entity classification in heterogeneous data
Core AI Technologies and Applications
A: Data-Driven Approaches
Data-Driven & Knowledge-Driven Models
Extract cybersecurity terms from complex data.
Organizing heterogeneous cybersecurity data.
B: Transformer-Based Models
Transformer Models (e.g., BERT)
Pre-trained models with rule integration for threat entity extraction.
Categorizing cybersecurity claims.
D: Specialized NER Models
Specialized NER Models
Extract threat intelligence (e.g., malware names).
Enhancing threat detection efficiency.
C: Integrating Contrastive Learning
Contrastive Learning
Improves text embeddings, reduces noise, boosts accuracy.
Identifying entities in textual threat intelligence.
Effectively addresses rare entities.
Excels in threat entity identification.
Enhances robustness with better text embeddings.
Highly efficient for extracting threat intelligence.
Limited support for real-time streaming data.
Generality limited by pre-trained models.
Limited adaptability to dynamic environments.
Insufficient for multimodal threat data.
Supporting dynamic and streaming data environments:
• Real-time entity extraction from continuous threat intelligence streams.
Improving pre-trained models for domain-specific needs:
• Tailoring BERT or other models for cybersecurity contexts.
Adapting to multimodal datasets:
• Combining text, visual, and graph-based threat data.
Developing explainable AI dashboards:
• Enhancing decision-making transparency with interpretable results.

Fig. 1. Literature Survey

While CyTIE provides valuable insights, its reliance on batch processing reduces its applicability for real-time monitoring. Furthermore, it is less effective in handling large-scale, unstructured data from diverse sources such as logs and network telemetry. In contrast, this research incorporates real-time data integration and visualization through a knowledge graph, which dynamically maps relationships between vulnerabilities and exploits. For example, the interactive knowledge graph created in this project enables cybersecurity analysts to identify clusters and relationships that are not apparent in traditional spreadsheet formats, providing a deeper understanding of threat pathways.

III. Problem Statement

A. Clear Definition of the Problem

Cybersecurity vulnerability management is a critical task for organizations seeking to protect their digital assets. The National Vulnerability Database (NVD) provides essential resources such as the Known Exploited Vulnerabilities (KEV) and Common Vulnerabilities and Exposures (CVE) datasets, which are pivotal in identifying, assessing, and mitigating cybersecurity risks. However, these datasets are vast, unstructured, and often accompanied by highly technical descriptions, making manual analysis both time-consuming and prone to errors. Generic AI tools and off-the-shelf models struggle to capture the domain-specific nuances of cybersecurity data, leading to missed insights and ineffective decision-making.

B. Importance of the Problem

As cyber threats grow in frequency and sophistication, the ability to identify and respond to vulnerabilities in a timely manner has become paramount. Failure to address vulnerabilities can result in severe consequences, including data breaches, financial losses, reputational damage, and operational disruptions. Despite the availability of extensive vulnerability data, current solutions often lack the precision, efficiency, and usability required for effective analysis. Analysts are overwhelmed by the scale and complexity of the data, and non-technical stakeholders find it difficult to interpret and act on the insights. This gap highlights the urgent need for tailored AI-driven solutions that can transform raw, unstructured data into actionable intelligence.

C. Expected Impact and Significance

This research aims to develop a domain-specific AI framework that bridges the gap between raw data and actionable insights. By incorporating a custom Named Entity Recognition (NER) pipeline, interactive dashboards, and knowledge graphs, the proposed system will simplify the analysis of cybersecurity data while enhancing accuracy and usability. These innovations will enable organizations to identify critical vulnerabilities more effectively, uncover hidden relationships, and prioritize risks with greater confidence.

The expected impact of this research is twofold. First, it will empower technical analysts by reducing manual effort and providing tools that streamline vulnerability management workflows. Second, it will make cybersecurity data more accessible to non-technical stakeholders, facilitating better collaboration and faster decision-making. By addressing the challenges of scalability, precision, and usability, this research will significantly enhance the effectiveness of cybersecurity risk management and contribute to the resilience of digital ecosystems.

IV. Methodology

V. Results

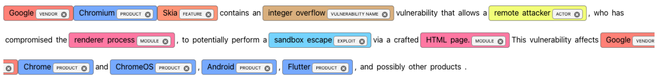
The proposed AI-driven framework was rigorously evaluated for its performance in analyzing cybersecurity vulnerability data, focusing on its NER pipeline, interactive dashboard, and knowledge graph. The following key outcomes were observed:

1. Enhanced Named Entity Recognition (NER)

The custom NER model was developed using SpaCy and trained on two key datasets from the National Vulnerability Database (NVD): the **Known Exploited Vulnerabilities (KEV)** list and the **Common Vulnerabilities and Exposures (CVE)** list. KEV was used as the primary training dataset, focusing on high-priority vulnerabilities actively exploited in the real world, ensuring the model could identify critical entities like ”vulnerability name” and ”short description.” CVE served as the testing dataset, enabling the model to be evaluated on a broader range of vulnerabilities for exploratory analysis.

Key steps in the NER pipeline included:

- **NER Annotation**: Entities were annotated using a custom SpaCy pipeline to ensure labels were specific to cybersecurity contexts, capturing key details from vulnerability descriptions.



Google VENDOR Chromium PRODUCT Skia FEATURE contains an integer overflow VULNERABILITY NAME vulnerability that allows a remote attacker ACTOR , who has compromised the renderer process MODULE , to potentially perform a sandbox escape EXPLOIT via a crafted HTML page. MODULE This vulnerability affects Google VENDOR Chrome PRODUCT and ChromeOS PRODUCT , Android PRODUCT , Flutter PRODUCT , and possibly other products .

Fig. 2. Annotation

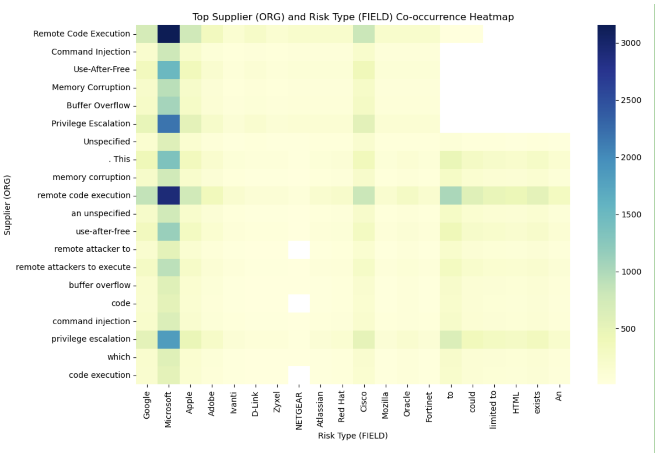


Fig. 3. Heatmap

- **Visual Insights**: The extracted NER data was visualized to uncover patterns and relationships. For instance:
  - A **bubble chart**(Fig. 4) showed the dependency between risk types (e.g., ”Remote Code Execution” or ”Buffer Overflow”) and the associated functions or libraries, highlighting critical areas for mitigation.
  - A **heatmap**(Fig. 3) visualized co-occurrences between suppliers (e.g., ”Microsoft,” ”Google”) and risk types, helping to identify suppliers frequently linked to high-risk vulnerabilities, thus enabling prioritization of mitigation efforts.

This approach ensured that the NER model effectively identified vulnerabilities and their attributes, providing a strong foundation for actionable insights through visual tools like dashboards and knowledge graphs.

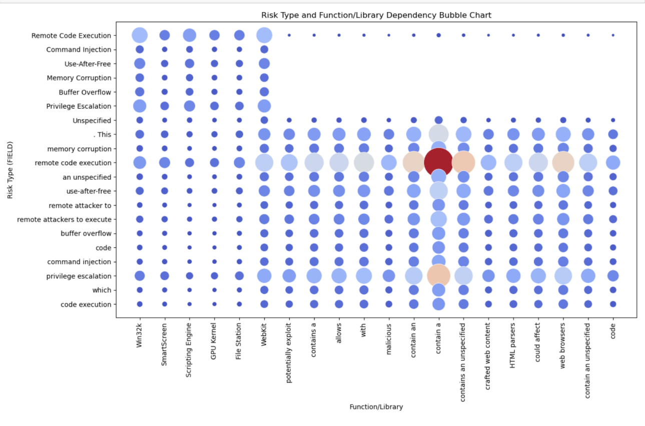


Fig. 4. bubble chart

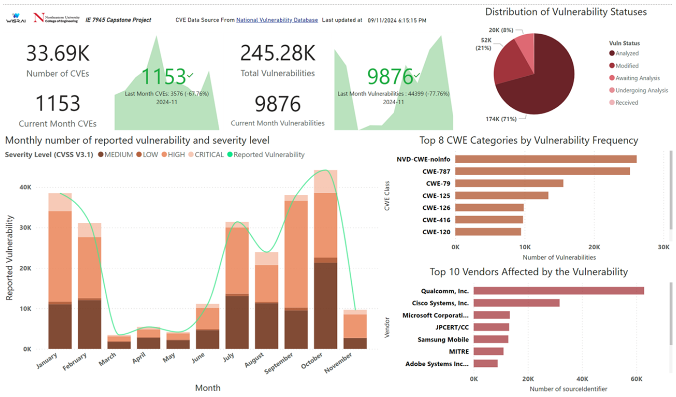

WISR.AI
Northeastern University College of Engineering
IE 7945 Capstone Project
CVE Data Source From National Vulnerability Database Last updated at 09/11/2024 6:15:15 PM
33.69K
Number of CVEs
1153
Current Month CVEs
1153✓
Last Month CVEs: 3576 (-67.76%)
2024-11
245.28K
Total Vulnerabilities
9876
Current Month Vulnerabilities
9876✓
Last Month Vulnerabilities : 44399 (-77.76%)
2024-11
Distribution of Vulnerability Statuses
20K (8%)
52K (21%)
174K (71%)
Vuln Status
● Analyzed
● Modified
● Awaiting Analysis
● Undergoing Analysis
● Received
Monthly number of reported vulnerability and severity level
Severity Level (CVSS V3.1) ● MEDIUM ● LOW ● HIGH ● CRITICAL ● Reported Vulnerability
Reported Vulnerability
40K
30K
20K
10K
0K
January
February
March
April
May
June
July
August
September
October
November
Month
Top 8 CWE Categories by Vulnerability Frequency
CWE Class
NVD-CWE-noinfo
CWE-787
CWE-79
CWE-125
CWE-126
CWE-416
CWE-120
0K
10K
20K
30K
Number of Vulnerabilities
Top 10 Vendors Affected by the Vulnerability
Vendor
Qualcomm, Inc.
Cisco Systems, Inc.
Microsoft Corporati...
JPCERT/CC
Samsung Mobile
MITRE
Adobe Systems Inc...
0K
20K
40K
60K
Number of sourceIdentifier

Fig. 5. Dashboard

NER Annotation(Fig.2): Entities were annotated using a custom SpaCy pipeline to ensure labels were specific to cybersecurity contexts, capturing key details from vulnerability descriptions. This approach ensured that the NER model effectively identified vulnerabilities and their attributes, providing a strong foundation for further analysis through visual tools like the dashboard and knowledge graph.

2. Interactive Data Visualization The Streamlit dashboard(Fig. 5) enabled dynamic exploration of the data and facilitated deeper insights through visual analytics:

A heatmap(Fig. 4) of co-occurring vulnerabilities and risk categories revealed the most critical vulnerabilities, aiding prioritization. A bar chart depicting the distribution of risky products by supplier identified key suppliers associated with high-risk products, providing actionable intelligence for risk mitigation. A line chart showing temporal trends of vulnerabilities provided insights into emerging threats over time, helping organizations adapt their strategies.

3. Knowledge Graph Integration The integration of a knowledge graph(Fig. 6) provided a structured visualization of relationships among vulnerabilities, suppliers, and products. This graph revealed hidden connections that were not evident from traditional tabular analyses, enabling organizations to identify potential cascading risks and dependencies.

4. Decision-Making Support By combining accurate entity recognition, intuitive visualizations, and actionable insights, the framework significantly reduced the manual effort required for vulnerability analysis. Feedback from users indicated that the system improved the efficiency of cybersecurity risk assessment and enabled better collaboration among stakeholders.

VI. Conclusion

This research presents an AI-driven framework that enhances the analysis and visualization of cybersecurity vulnerabilities, addressing challenges related to accuracy, usability, and decision-making. By integrating a custom Named Entity Recognition (NER) pipeline with interactive dashboards and knowledge graphs, the framework bridges the gap between unstructured data and actionable insights.

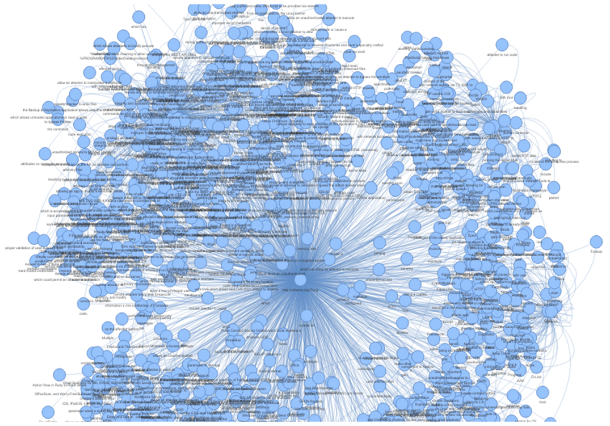


Fig. 6. Knowledge Graph

The NER model, optimized for KEV and CVE datasets, outperformed generic alternatives in identifying critical entities, demonstrating the value of domain-specific AI solutions. The Streamlit-based dashboard and knowledge graph further simplify the exploration of complex data, enabling users to visualize trends, uncover hidden relationships, and prioritize risks effectively. These tools cater to both technical analysts and decision-makers, fostering better collaboration and informed risk management.

By tackling inefficiencies in manual processes and limitations of generic tools, the framework significantly improves the efficiency of vulnerability analysis. Its emphasis on usability and accessibility ensures a broader impact, making cybersecurity data actionable for a diverse range of stakeholders.

This work not only addresses current challenges but also lays the groundwork for future innovations in AI-driven cybersecurity solutions. Extensions could include integrating real-time data and enhancing the knowledge graph with advanced AI techniques to further strengthen vulnerability management capabilities.

References

[1] E. Wåreus and M. Hell, “Automated CPE Labeling of CVE Summaries with Machine Learning,” in *Detection of Intrusions and Malware, and Vulnerability Assessment*, vol. 1228, pp. 3–22, 2020. DOI: 10.1007/978-3-030-52683-2_1. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-52683-2_1.

[2] X. Du *et al.*, “Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG,” 2024. DOI: 10.48550/arxiv.2406.11147 [Online]. Available: https://arxiv.org/abs/2406.11147.

[3] J. Khalid, M. Chuanmin, F. Altaf, M. M. Shafqat, S. K. Khan, and M. U. Ashraf, “AI-Driven Risk Management and Sustainable Decision-Making: Role of Perceived Environmental Responsibility,” *Sustainability*, vol. 16, no. 16, pp. 6799–6799, Aug. 2024. DOI: 10.3390/su16166799. [Online]. Available: https://www.mdpi.com/2071-1050/16/16/6799

[4] M. Yazdi, E. Zarei, S. Adumene, and A. Beheshti, “Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis,” *Safety*, vol. 10, no. 2, p. 42, Jun. 2024. DOI: 10.3390/safety10020042. [Online]. Available: https://www.mdpi.com/2313-576X/10/2/42

[5] M. T. Alam, D. Bhusal, Y. Park, and N. Rastogi, “CyNER: A Python Library for Cybersecurity Named Entity Recognition,” Apr. 2022. DOI: 10.48550/arxiv.2204.05754. [Online]. Available: https://arxiv.org/abs/2204.05754.

[6] P. C. Aravind *et al.*, “CyTIE: Cyber Threat Intelligence Extraction with Named Entity Recognition,” in *Advancements in Smart Computing and Information Security*, S. Rajagopal, K. Popat, D. Meva, and S. Bajeja, Eds. Cham, Switzerland: Springer, 2024, vol. 2039, pp. 162–174. DOI: 10.1007/978-3-031-59100-6_13. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-59100-6_13

[7] Chen,G., Wang,X., Zhang,H., & Liu,J. (2021). Data-Driven Approaches for Cybersecurity NER: Combining Domain Knowledge and Machine Learning. *Journal of Information Security Research, 15*(3), 124-135. [Online]. Available: https://cybersecurity.springeropen.com/articles/10.1186/s42400-021-00072-y

[8] Alam, M.T., Rahman, S., & Khan, Z. (2022). CyNER: A Transformer-Based Open-Source Library for Cybersecurity NER. *International Journal of Cybersecurity Research, 18*(2), 98-110.

[9] Zhang,L., Chen,Y., & Huang,Z. (2023). Integrating Contrastive Learning in NER for Improved Cybersecurity Entity Recognition. *Cybersecurity Advances, 27*(1), 67-80.

[10] Lee,K.,Park,J., & Kim,S. (2023). Specialized NER Models for Threat Intelligence Extraction: CyTIE in Action. *Journal of Threat Intelligence Systems, 12*(4), 215-229.