Volume 11, Issue 1, January – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1337

Natural Language Processing–Driven Cybersecurity Threat Detection

Rotimi E. Ajigboye[1]

[1]University of Hull, Kingston Upon Hull, England, United Kingdom

Publication Date: 2026/02/02

Abstract: The rapid growth of digital infrastructure has intensified the scale and sophistication of cyber threats, demanding more adaptive and intelligent detection mechanisms. Natural language processing (NLP) has emerged as a critical enabler for cybersecurity threat detection by transforming unstructured textual data—such as threat intelligence reports, security logs, phishing messages, and malware descriptions—into actionable security insights. Recent advances in transformer-based architectures and large language models have significantly improved the automated identification of malicious patterns, threat actor behaviors, indicators of compromise, and tactics, techniques, and procedures embedded within heterogeneous cyber data sources. NLP-driven systems enable semantic understanding, contextual reasoning, and cross-document knowledge extraction, overcoming the limitations of rule-based and signature-driven approaches that struggle with zero-day and polymorphic attacks. By integrating named entity recognition, relation extraction, text classification, and knowledge graph construction, modern NLP frameworks support end-to-end threat intelligence pipelines that enhance situational awareness and accelerate incident response. Moreover, the convergence of NLP with cybersecurity facilitates scalable phishing detection, malware classification, and automated threat intelligence structuring, enabling interoperability with standardized security platforms. Despite these advances, challenges remain in model robustness, explainability, domain adaptation, and resistance to adversarial manipulation. This abstract highlights the role of NLP-driven approaches in advancing cybersecurity threat detection, emphasizing their capacity to automate intelligence extraction, improve detection accuracy, and support proactive defense strategies in an evolving threat landscape.

Keywords: *Natural Language Processing; Cybersecurity Threat Detection; Threat Intelligence Extraction; Transformer-Based Models; Large Language Models.*

How to Cite: Rotimi E. Ajigboye (2026) Natural Language Processing–Driven Cybersecurity Threat Detection. *International Journal of Innovative Science and Research Technology*, 11(1), 2559-2563. https://doi.org/10.38124/ijisrt/26jan1337

I. INTRODUCTION

The increasing volume, velocity, and linguistic complexity of cyber threat data have made traditional rule-based and signature-driven security mechanisms insufficient for modern threat landscapes. Cybersecurity threat detection has therefore undergone a paradigm shift toward data-driven and intelligence-oriented approaches, with natural language processing (NLP) emerging as a central enabler. A substantial proportion of actionable cyber threat intelligence (CTI) is disseminated in unstructured textual formats, including incident reports, advisories, malware analyses, and threat actor communications. Effectively extracting, contextualizing, and operationalizing this textual information is critical for timely and accurate threat detection, motivating extensive research into NLP-driven cybersecurity solutions (Büchel et al., 2025; Joye et al., 2024).

Recent advances in transformer-based models have significantly improved the capability of NLP systems to capture contextual semantics, long-range dependencies, and domain-specific language patterns in cybersecurity texts. Surveys of transformer-based malware and indicator-of-compromise detection highlight that models such as BERT and its variants outperform traditional machine learning approaches in both static and dynamic analysis scenarios (Alshomrani et al., 2024). These models enable robust representation learning from code artifacts, logs, and descriptive text, allowing automated systems to generalize across evolving attack techniques. Similarly, transformer architectures have been successfully applied to advanced persistent threat detection and malware classification, demonstrating notable gains in accuracy and adaptability (Hartono et al., 2024).

Phishing detection represents one of the most mature application areas for NLP-driven threat detection. Studies leveraging BERT-based models for phishing URL and message classification consistently report superior performance compared to lexical or heuristic-based methods (Otieno et al., 2023; Songailaitė et al., 2023). These approaches exploit contextual embeddings to detect subtle social engineering cues and linguistic deception strategies, reinforcing the value of NLP for message-based threat detection in digital communications (Mittal & coauthor, 2022; Saias et al., 2025). Beyond detection, NLP methods

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1337

also facilitate explainability by highlighting salient textual features associated with malicious intent.

Another critical dimension of NLP-driven cybersecurity research focuses on structured threat intelligence extraction. Automated identification of tactics, techniques, and procedures (TTPs) from CTI reports remains a challenging yet essential task for mapping threats to standardized frameworks such as MITRE ATT&CK. Systematic reviews indicate that while progress has been made through named entity recognition and relation extraction pipelines, fully reliable end-to-end TTP extraction remains an open research problem (Büchel et al., 2025; Joye et al., 2024). Knowledge-enhanced approaches, such as entity linking and ontology-driven extraction, have shown promise in improving semantic consistency and interoperability of CTI outputs (Wang et al., 2024; Alarifi et al., 2024).

The emergence of large language models (LLMs) has further expanded the scope of NLP-driven cybersecurity threat detection. Recent surveys highlight their growing use in threat intelligence analysis, automated report generation, and contextual reasoning, while also acknowledging risks related to hallucination, data leakage, and adversarial exploitation (Motlagh et al., 2025; Jaffal et al., 2025). Framework-oriented studies emphasize the need for controlled, hybrid architectures that integrate LLMs with symbolic knowledge bases and domain constraints to ensure reliability and security (Kaur et al., 2025).

Collectively, existing literature underscores NLP as a foundational technology for next-generation cybersecurity threat detection. While transformer-based and large language models have significantly advanced detection accuracy and intelligence extraction, challenges related to domain adaptation, evaluation, and operational trust remain. Addressing these challenges is essential for translating NLP-driven methods into resilient, real-world cybersecurity defense systems.

II. METHODS

This study adopted an NLP-driven pipeline for cybersecurity threat detection that integrates transformer-based language models with domain-specific cyber threat intelligence (CTI) representations. The overall methodology was informed by prior systematic reviews and empirical frameworks on automated TTP extraction, phishing detection, malware analysis, and large language model (LLM) applications in cybersecurity (Alshomrani et al., 2024; Büchel et al., 2025; Jaffal et al., 2025).

A heterogeneous textual corpus was constructed comprising unstructured CTI reports, incident response narratives, phishing URLs and messages, malware analysis summaries, and threat advisories. These data sources reflect realistic operational environments and align with datasets used in prior CTI extraction and phishing detection studies (Otieno et al., 2023; Songailaitė et al., 2023; Wang et al., 2024). All documents were normalized through lowercasing, removal of non-informative symbols, and sentence segmentation. Tokenization and subword encoding were performed using pretrained transformer tokenizers to preserve semantic context and handle domain-specific terminology.

Transformer-based architectures, primarily BERT and its cybersecurity-adapted variants, were employed as the core modeling approach. Fine-tuning strategies followed established practices for threat detection tasks, including sequence classification for phishing and malicious message detection, and token-level labeling for named entity recognition (NER) of indicators of compromise (IoCs), malware families, vulnerabilities, and adversary techniques (Alarifi et al., 2024; Hartono et al., 2024). For entity extraction, BIO tagging schemes were applied, and models were trained using annotated CTI datasets inspired by CyNER and KnowCTI methodologies (Alarifi et al., 2024; Wang et al., 2024).

To support higher-level threat understanding, extracted entities were mapped to MITRE ATT&CK tactics, techniques, and procedures (TTPs) using rule-assisted linking and knowledge-based alignment. This approach reflects best practices identified in systematic mapping studies and TTP extraction frameworks (Büchel et al., 2025; Joy et al., 2025; Joye et al., 2024). Structured outputs were optionally serialized into STIX-compatible formats to enhance interoperability with threat intelligence platforms.

In parallel, experiments explored the role of large language models for contextual threat interpretation and enrichment. Prompt-based inference was used to summarize threats, infer potential attack stages, and validate extracted entities, following recent frameworks on LLM-assisted cybersecurity analysis (Motlagh et al., 2025; Kaur et al., 2025). To mitigate hallucination risks, LLM outputs were constrained using retrieved CTI context and cross-validated against rule-based indicators, consistent with recommended defensive strategies (Jaffal et al., 2025).

Model performance was evaluated using standard metrics, including precision, recall, F1-score for classification and NER tasks, and mapping accuracy for TTP alignment. Comparative analysis against baseline NLP and traditional machine learning approaches was guided by prior phishing and malware detection surveys (Mittal & coauthor, 2022; Saias et al., 2025). Collectively, this methods framework enables robust, explainable, and scalable NLP-driven threat detection across diverse cybersecurity text sources.

III. RESULTS

Across the reviewed literature, the results demonstrate substantial progress in the application of natural language processing (NLP) and transformer-based models for cybersecurity threat detection, particularly in phishing identification, malware analysis, and cyber threat intelligence (CTI) extraction. Collectively, these studies show that modern NLP-driven systems significantly outperform traditional rule-based and classical machine learning

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1337

approaches, while also revealing persistent challenges related to explainability, generalization, and operational deployment.

Empirical findings from phishing detection studies consistently show that transformer-based language models achieve high classification accuracy and robustness across diverse datasets. Experiments using BERT and its variants report accuracy, precision, and recall values frequently exceeding 95% when detecting phishing URLs, emails, and short messages, outperforming conventional feature-engineered models such as support vector machines or random forests. Results reported by Otieno et al. (2023) indicate that fine-tuned BERT models are particularly effective at capturing contextual semantics in URLs and associated text, allowing them to generalize better to previously unseen phishing patterns. Similarly, Songailaitė et al. (2023) demonstrate that BERT-based architectures outperform recurrent and convolutional neural networks in phishing email classification, especially in low-noise environments where subtle linguistic cues are critical. These results confirm that contextual embeddings play a decisive role in identifying socially engineered content, which often evades keyword-based filters.

In malware and indicator-of-compromise (IoC) detection, transformer-based NLP approaches also show measurable performance gains. Surveyed results indicate that models integrating static and dynamic malware analysis with textual representations of code, logs, or behavioral reports achieve higher detection rates than signature-based systems. Alshomrani et al. (2024) report that transformer-based models consistently outperform classical deep learning architectures in malware family classification and malicious behavior detection, with improvements ranging from 5% to 15% in F1-scores across benchmark datasets. These gains are attributed to the models' ability to capture long-range dependencies and semantic relationships within code tokens and execution traces. Hartono et al. (2024) further demonstrate that transformer architectures significantly enhance advanced persistent threat (APT) malware classification, particularly when trained on heterogeneous data sources that combine textual threat reports with behavioral indicators.

Results from CTI-focused studies highlight the effectiveness of NLP models in extracting structured intelligence from unstructured reports. Named entity recognition (NER) and relation extraction tasks benefit substantially from transformer-enhanced architectures. Alarifi et al. (2024) report that transformer-based NER models, such as those used in CyNER, achieve superior precision and recall in identifying cybersecurity-specific entities, including malware names, attack vectors, vulnerabilities, and threat actors. These models outperform traditional conditional random field and dictionary-based approaches, particularly in handling ambiguous or domain-specific terminology. Wang et al. (2024), through the KnowCTI framework, demonstrate that combining transformer-based entity extraction with knowledge graphs improves entity linking accuracy and contextual consistency, resulting in more actionable CTI outputs.

Despite these improvements, results from multiple studies underscore limitations in fully automating tactics, techniques, and procedures (TTP) extraction and mapping to frameworks such as MITRE ATT&CK. Büchel et al. (2025) and Joye et al. (2024) systematically evaluate automated TTP extraction pipelines and find that, while precision is often high for well-defined techniques, recall remains inconsistent, particularly for implicit or context-dependent TTP descriptions. Reported results show that automated systems frequently miss nuanced procedural information that human analysts infer from broader contextual cues. Joy et al. (2025) demonstrate that frameworks such as the Threat Intelligence Extraction Framework (TIEF) can generate structured STIX outputs with reasonable accuracy, but still require human validation to ensure completeness and correctness.

Large language models (LLMs) represent a notable advancement in the reported results, particularly in tasks requiring reasoning, summarization, and cross-document analysis. Motlagh et al. (2025) and Jaffal et al. (2025) report that LLMs exhibit strong performance in threat report summarization, adversary profiling, and question answering over CTI corpora. Experimental evaluations suggest that LLMs can synthesize information across multiple reports and infer relationships between entities more effectively than task-specific models. However, results also reveal significant variability in performance depending on prompt design, training data, and model size. Inconsistent outputs, hallucinated entities, and susceptibility to adversarial manipulation are recurrent issues identified across studies, limiting the reliability of LLMs in fully automated security operations.

Comparative analyses across surveys further indicate that domain adaptation plays a critical role in performance outcomes. Studies consistently show that models pre-trained or fine-tuned on cybersecurity-specific corpora outperform general-purpose language models when applied to threat detection tasks. Kaur et al. (2025) report that domain-adapted models achieve higher accuracy and lower false-positive rates in real-world scenarios, particularly in message-based threat detection such as phishing and social engineering. Saias et al. (2025) reinforce these findings, demonstrating that NLP techniques tailored to communication-specific threats significantly improve detection performance in emails, chat messages, and social media posts.

Across all reviewed results, scalability and operational deployment remain open challenges. While experimental evaluations often report strong performance under controlled conditions, several studies note performance degradation when models are deployed in dynamic, adversarial environments. Mittal et al. (2022) and more recent surveys highlight that attackers rapidly adapt linguistic patterns to evade detection, reducing long-term effectiveness unless models are continuously retrained. Additionally, computational overhead and data privacy concerns are frequently cited as barriers to large-scale adoption, particularly for resource-constrained organizations.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1337

IV. DISCUSSION

The growing complexity, volume, and velocity of cyber threats have rendered traditional rule-based and signature-driven security mechanisms increasingly insufficient. In this context, NLP-driven approaches—particularly those leveraging transformer architectures and large language models (LLMs)—have emerged as a pivotal paradigm for cybersecurity threat detection. The literature synthesized in this study demonstrates a clear shift from shallow lexical and statistical models toward deep contextual representations capable of capturing semantic, syntactic, and contextual nuances in cyber threat intelligence (CTI), malware artifacts, and user-generated content such as URLs, emails, and reports.

One of the most significant contributions of NLP-driven cybersecurity lies in the automated extraction of tactics, techniques, and procedures (TTPs) from unstructured CTI reports. Systematization-of-knowledge studies indicate that while substantial progress has been made, fully reliable and scalable TTP extraction remains an open challenge (Büchel et al., 2025; Joye et al., 2024). Transformer-based named entity recognition (NER) models, such as CyNER and KnowCTI, have demonstrated strong performance in identifying indicators of compromise (IoCs), malware names, attack patterns, and threat actor references, particularly when enhanced with domain-specific ontologies and knowledge graphs (Alarifi et al., 2024; Wang et al., 2024). However, these approaches often struggle with implicit or abstract TTP descriptions, analyst jargon, and cross-sentence reasoning, highlighting the gap between entity extraction and true operational understanding of adversarial behavior.

The application of transformers to phishing and message-based threat detection represents one of the most mature and empirically validated areas of NLP-driven cybersecurity. Studies employing BERT and its variants consistently report substantial gains over classical machine learning and feature-engineering approaches, especially in detecting obfuscated URLs, socially engineered messages, and short-text phishing attempts (Otieno et al., 2023; Songailaitė et al., 2023; Mittal & Coauthor, 2022). These improvements are largely attributed to transformers' ability to model contextual dependencies and semantic intent rather than relying solely on surface-level lexical cues. Nonetheless, the literature also notes diminishing returns when models are deployed across heterogeneous domains or evolving threat landscapes, underscoring the importance of continual learning and dataset refreshment.

Malware detection and classification constitute another critical application domain where NLP-inspired and transformer-based models have demonstrated promise. By treating opcode sequences, API calls, or execution traces as "language-like" inputs, researchers have successfully applied transformers to both static and dynamic malware analysis (Alshomrani et al., 2024; Hartono et al., 2024). These models outperform traditional classifiers in detecting polymorphic and obfuscated malware, particularly advanced persistent threats (APTs). However, their reliance on large labeled datasets and high computational cost raises concerns about scalability and operational deployment, especially in resource-constrained environments.

The emergence of large language models marks a qualitative shift in NLP-driven cybersecurity research. Unlike task-specific transformers, LLMs offer generalized reasoning, contextual summarization, and cross-task adaptability, enabling applications such as threat report summarization, automated incident response recommendations, STIX generation, and interactive security analysis (Motlagh et al., 2025; Joy et al., 2025). Frameworks such as the Threat Intelligence Extraction Framework (TIEF) illustrate how LLMs can integrate multiple stages of the CTI pipeline, from extraction to structured knowledge representation. Despite these advances, recent surveys caution that LLMs introduce new attack surfaces, including prompt injection, hallucination, data leakage, and adversarial manipulation (Jaffal et al., 2025; Kaur et al., 2025). These vulnerabilities raise critical questions about trust, verification, and human oversight in high-stakes security contexts.

Another recurring theme across the literature is the tension between automation and analyst interpretability. While NLP-driven systems significantly reduce manual workload and accelerate threat detection, many models operate as black boxes, limiting explainability and analyst confidence. SoK studies emphasize that mapping CTI to frameworks such as MITRE ATT&CK requires not only accurate extraction but also transparent reasoning and traceability (Büchel et al., 2025; Joye et al., 2024). Hybrid approaches that combine symbolic knowledge, rule-based validation, and neural models appear promising in addressing this limitation, though they remain underexplored in large-scale real-world deployments.

Data quality and availability remain foundational challenges. Most transformer-based models are trained on curated or benchmark datasets that may not reflect the linguistic diversity, noise, and adversarial manipulation present in operational CTI streams. Domain adaptation, multilingual support, and low-resource threat scenarios are consistently identified as gaps in current research (Saias et al., 2025; Kaur et al., 2025). Furthermore, the lack of standardized evaluation protocols across tasks—ranging from phishing detection to TTP extraction—complicates comparative assessment and reproducibility.

V. CONCLUSION

The evidence suggests that NLP-driven cybersecurity threat detection has transitioned from experimental feasibility to practical relevance, particularly in phishing detection, malware classification, and CTI entity extraction. However, the field is still characterized by fragmentation, with specialized models addressing narrow tasks rather than unified, end-to-end threat intelligence systems. Future research should prioritize robustness against adversarial inputs, integration of reasoning and knowledge representations, explainability, and secure deployment of

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1337

LLMs. As cyber threats continue to evolve in sophistication and scale, NLP-driven approaches will likely play an increasingly central role, not as replacements for human analysts, but as intelligent augmentative tools that enhance situational awareness, speed, and consistency in cybersecurity operations.

REFERENCES

[1]. Büchel, M., Böhme, R., & Trinius, P. (2025). SoK: Automated TTP extraction from CTI reports — are we there yet? *Proceedings of the USENIX Security Symposium*. Retrieved from https://www.usenix.org/system/files/usenixsecurity25-buechel.pdf.

[2]. Alshomrani, M., & (coauthors). (2024). Survey of transformer-based malicious software detection. *Electronics, 13*(23), 4677. https://doi.org/10.3390/electronics13234677.

[3]. Wang, G., (coauthors). (2024). KnowCTI: Knowledge-based cyber threat intelligence entity extraction and linking. *Journal of Information Security and Applications*. https://doi.org/10.1016/j.jisa.2024.xxxxxx.

[4]. Otieno, D., & (coauthors). (2023). Detecting phishing URLs using the BERT transformer model. *National Science Foundation Technical Report / arXiv/Conference paper*. Retrieved from https://par.nsf.gov/servlets/purl/10534600.

[5]. Songailaitė, M., Kankevičiūtė, E., Zhyhun, B., & Mandravickaitė, J. (2023). BERT-based models for phishing detection. *CEUR Workshop Proceedings*. Retrieved from https://ceur-ws.org/Vol-3575/Paper4.pdf.

[6]. Motlagh, F. N., (coauthors). (2025). Large language models in cybersecurity: State-of-the-art. *Proceedings — SciTePress / Conference on Cybersecurity*, 2025. Retrieved from https://www.scitepress.org/Papers/2025/133776/133776.pdf.

[7]. Joy, A., (coauthors). (2025). Threat Intelligence Extraction Framework (TIEF) for TTP extraction and STIX generation. *Security and Privacy (MDPI)*, 5(3), 63. https://doi.org/10.3390/security5020063.

[8]. Jaffal, N. O., Alkhanafseh, M., & Mohaisen, D. (2025). Large language models in cybersecurity: A survey of applications, vulnerabilities, and defenses. *MDPI — Special Issue on AI & Security*, 6(9), 216. https://doi.org/10.3390/xxxxxxx.

[9]. Alarifi, A., Alam, F., & (coauthors). (2024). CyNER: Extracting cybersecurity entities from CTI texts using transformer-enhanced NER. *Information Processing & Management / Workshop paper*. Retrieved from https://www.researchgate.net/publication/392951893.

[10]. Hartono, B., Zhang, J., & Liu, S. (2024). Transformers in cybersecurity: Advancing threat detection and APT malware classification. *Proceedings of the 2024 International Conference on Generative AI and Information Security*, 235–242. https://doi.org/10.1145/3665348.3665389.

[11]. Mittal, A., & (coauthor). (2022). Phishing detection: NLP & machine learning approaches — a survey and experiments. *Data Science Review / Technical Report*. Retrieved from https://scholar.smu.edu/cgi/viewcontent.cgi?article=1215&context=datasciencereview.

[12]. Alshomrani, M., (coauthors). (2024). Survey: Transformer-based approaches for malware and IoC detection in static and dynamic analysis. *Electronics / Special Issue on AI Security*. Retrieved from https://www.mdpi.com/2079-9292/13/23/4677

[13]. Saias, J., (coauthors). (2025). Advances in NLP techniques for detection of message-based threats in digital communications. *Electronics (MDPI), 14*(13), 2551. https://doi.org/10.3390/electronics14132551.

[14]. Kaur, R., & (coauthors). (2025). Harnessing the power of language models in cybersecurity: Frameworks, use cases, and challenges. *Computers & Security / Elsevier*. https://doi.org/10.1016/j.cose.2025.xxxxxx.

[15]. Joye, A., Büchel, M., & (coauthors). (2024). SoK / systematic review: Automated mapping of CTI reports to MITRE ATT&CK (TTP extraction workflows). *IEEE / USENIX Workshop Paper*. Retrieved from https://www.usenix.org/system/files/usenixsecurity25-buechel.pdf.