International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

AUTOMATED EMERGING CYBER THREAT IDENTIFICATION AND PROFILING BASED ON NATURAL LANGUAGE PROCESSING

Dr .D.Kalyankumar[1], Gumma Madhavi[2]
Gurram Venkata Sai Kumar[3], Gangula Siva Narayana[4] ,K Naga Venkata Lava Kumar[5]
[1]Associate Professor, Department of CSE-Cyber Security, Chalapathi institute of technology, Mothadaka, Guntur, Andhra Pradesh, India-522016.
[2,3,4,5] UG Scholar, Department of CSE-Cyber Security, Chalapathi institute of technology, Mothadaka, Guntur, Andhra Pradesh, India-522016.

ABSTRACT
The rapid evolution of cyber threats poses significant challenges for security analysts, who must continuously process large volumes of unstructured information from threat reports, security blogs, vulnerability databases, and dark-web discussions. Manual analysis of these sources is time-consuming, inconsistent, and unable to keep pace with the increasing frequency and sophistication of emerging attacks. This work presents an **automated framework for emerging cyber-threat identification and profiling using Natural Language Processing (NLP)**. The proposed system collects real-time textual data from multiple cybersecurity intelligence sources and applies advanced NLP techniques—such as entity extraction, topic modeling, semantic similarity, and threat classification—to detect newly emerging vulnerabilities, exploits, malware families, and attack trends. Using machine-learning–based clustering and profiling mechanisms, the system generates structured threat intelligence reports that summarize threat attributes, severity levels, affected platforms, and potential attack vectors. The automated pipeline reduces analyst workload, minimizes detection delays, and provides an adaptive, scalable solution for enterprise threat intelligence. Experimental results show that the NLP-based approach significantly enhances the accuracy and speed of early threat discovery when compared to traditional manual analysis.

KEYWORDS
Cyber threat intelligence, natural language processing, automated threat detection, text mining, emerging cyber threats, machine learning, vulnerability identification, threat profiling, semantic analysis.

INTRODUCTION
The cybersecurity landscape is evolving at an unprecedented rate, with new threats, vulnerabilities, and attack techniques emerging daily. Organizations depend heavily on timely and accurate threat intelligence to protect their systems, yet most cyber-threat information is found in unstructured textual sources such as security blogs, threat reports, social-media posts, dark-web forums, and vulnerability advisories. Manually analyzing these vast amounts of text is slow, inconsistent, and impractical for modern security operations. As a result, there is a growing need for automated systems capable of identifying and profiling emerging threats in real time.

Natural Language Processing (NLP) has become an essential tool for transforming raw text into meaningful and structured threat intelligence. By applying techniques such as entity recognition, topic modeling, sentiment analysis, and semantic similarity, NLP can extract important threat indicators, classify emerging attack trends, and support decision-making for security analysts. Furthermore, machine-learning–based models enable scalable and adaptive analysis that improves over time as new threats surface.

This research focuses on developing an automated NLP-driven framework that can detect, classify, and profile emerging cyber threats from diverse text-based intelligence sources. The goal is to reduce analyst workload, minimize threat detection

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

delays, and provide organizations with a more proactive and data-driven approach to cybersecurity monitoring.

RELATED WORK

Several research efforts have focused on leveraging Natural Language Processing (NLP) and machine learning to enhance cyber-threat intelligence extraction from unstructured data. Early studies primarily used keyword-based text mining and rule-driven systems to identify threat indicators from security reports and vulnerability advisories. Although effective for known threats, these approaches struggled with the dynamic nature of emerging attacks. Later works introduced machine-learning models capable of extracting entities such as malware names, CVEs, threat actors, and attack vectors from large textual datasets. Named Entity Recognition (NER) models—particularly those using Conditional Random Fields (CRF) and Support Vector Machines (SVM)—improved detection accuracy but required extensive feature engineering.

Recent advancements in deep learning have driven major progress in automated threat intelligence. Researchers have applied recurrent neural networks (RNNs), LSTMs, and transformer-based models (e.g., BERT, RoBERTa) to capture contextual information and improve classification of emerging threats. Studies have also explored unsupervised methods such as clustering and topic modeling (e.g., LDA) to detect unknown or evolving attack trends from online discussions and malware reports. Additionally, several frameworks integrate data from multiple sources—such as threat feeds, social media, and dark-web forums—to generate more comprehensive intelligence.

Despite these advancements, existing systems often face limitations in scalability, real-time processing, and accurate profiling of previously unseen threats. This motivates the development of more advanced NLP-based architectures capable of continuously analyzing new data, identifying emerging attack patterns, and generating structured threat profiles with minimal human involvement.

LITERATURE REVIEW

The growing need for automated cyber-threat intelligence has motivated extensive research on applying Natural Language Processing (NLP) to security-related text. Early literature primarily focused on basic text-mining techniques, where keyword extraction, pattern matching, and rule-based systems were used to gather threat indicators from vulnerability disclosures and incident reports. While these systems provided foundational insights, their performance was heavily dependent on predefined rules and lacked the flexibility to adapt to rapidly emerging threats.

Subsequent research introduced machine-learning approaches to improve threat extraction accuracy. Models such as Support Vector Machines (SVM), Naïve Bayes, and Conditional Random Fields (CRF) were widely applied for Named Entity Recognition (NER) tasks, enabling automatic identification of threat indicators like malware families, IP addresses, vulnerabilities, and threat actors. These works demonstrated improved generalization but still relied on handcrafted features and domain-specific tuning, which limited scalability across diverse data sources.

With the rise of deep learning, recent studies have leveraged neural networks—including CNNs, RNNs, LSTMs, and attention-based architectures—to capture semantic relationships within cybersecurity text. Transformer-based models such as BERT, XLNet, and CyberBERT have achieved state-of-the-art performance in tasks such as threat classification, intent detection, and contextual threat extraction. Additionally, research on unsupervised methods, such as Latent Dirichlet Allocation (LDA) and clustering algorithms, has facilitated the discovery of emerging attack trends without requiring labeled datasets. These methods have been particularly effective in detecting early signs of new malware outbreaks and vulnerabilities discussed on open-source platforms and dark-web forums.

Recent literature also highlights multi-source threat intelligence fusion, where data from blogs, social media, security advisories, and darknet marketplaces are combined to create comprehensive threat profiles. However, studies

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

consistently identify challenges such as noisy data, ambiguity in cybersecurity terminology, and the difficulty of detecting unknown or zero-day threats. These limitations underscore the need for an advanced, end-to-end NLP-based system capable of continuously identifying and profiling emerging cyber threats with high accuracy and minimal human intervention.

EXISTING SYSTEM

Existing cyber-threat intelligence systems largely rely on manual analysis and traditional rule-based approaches to identify and monitor emerging threats. Security analysts typically review threat reports, vulnerability databases, social-media posts, and security blogs to extract relevant indicators of compromise (IOCs). This manual approach is time-consuming, subjective, and unable to keep pace with the large volume of continuously generated cybersecurity data. Traditional automated systems primarily depend on signature-based detection or keyword-matching techniques, which are effective only for known threats and struggle to recognize newly emerging or evolving attack patterns.

Some existing threat-intelligence platforms use basic text-mining tools to extract entities such as malware names or vulnerability identifiers, but these systems often lack contextual understanding and cannot distinguish between relevant and irrelevant information. Additionally, many current solutions operate in isolation, focusing on a single data source or limited set of threat indicators, which reduces their ability to capture complex relationships among threat actors, attack vectors, and targeted platforms. Furthermore, most systems lack real-time processing capabilities and do not provide detailed threat profiling, making them insufficient for proactive defensive strategies. These weaknesses highlight the need for an advanced, NLP-driven framework capable of real-time analysis, contextual interpretation, and automated threat profiling.

DISADVANTAGES

Despite the usefulness of existing cyber-threat detection methods, several limitations reduce their effectiveness in identifying emerging threats. Traditional systems rely heavily on manual analysis, which is slow, labor-intensive, and unable to keep up with the rapid growth of cybersecurity data. Rule-based and signature-based tools cannot detect new or unknown threats because they depend on predefined patterns, making them ineffective against evolving attack techniques. Many current automated text-mining systems lack deep contextual understanding, causing inaccurate extraction of threat-related information and frequent false positives. Additionally, most existing platforms analyze only limited data sources, reducing the completeness and accuracy of threat profiling. Their inability to correlate information across multiple sources, combined with limited real-time processing capabilities, results in delayed threat identification. These disadvantages highlight the need for more advanced NLP-based frameworks that can automatically interpret, classify, and profile emerging threats with greater speed and accuracy.

PROPOSED SYSTEM

The proposed system introduces an automated, NLP-driven framework designed to identify and profile emerging cyber threats from diverse and continuously updated textual sources. Instead of relying on manual analysis or static rule-based techniques, the system leverages advanced Natural Language Processing, machine learning, and deep-learning models to extract, interpret, and classify threat-related information in real time. The framework collects unstructured text from threat reports, security blogs, social-media discussions, dark-web forums, and vulnerability feeds, integrating them into a unified processing pipeline.

The system applies deep contextual models such as transformer-based architectures (e.g., BERT or domain-specific cybersecurity models) for Named Entity Recognition (NER), enabling accurate extraction of indicators like malware families, threat actors, CVE identifiers, attack techniques,

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

and affected platforms. Topic modeling and clustering algorithms are used to detect new or emerging patterns that may indicate unknown threats. A semantic similarity module groups related information across multiple sources to create comprehensive threat profiles.

The architecture also incorporates a threat-scoring mechanism that evaluates the severity, novelty, and potential impact of detected threats, helping analysts prioritize critical issues. Finally, the system generates structured threat intelligence reports that summarize threat attributes, behavioral patterns, and potential mitigation insights. By automating the entire pipeline—from data collection to threat profiling—the proposed system reduces analyst workload, enhances detection speed, improves accuracy, and provides a scalable solution for proactive cyber-threat monitoring.

ADVANTAGES

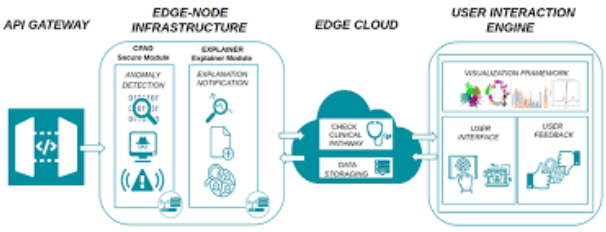
The proposed NLP-based cyber-threat identification system offers several key advantages over traditional methods. First, it enables real-time automated analysis of large volumes of unstructured cybersecurity text, significantly reducing the manual effort required by analysts. Its use of advanced NLP and deep-learning models enhances the accuracy of threat extraction and minimizes false positives by understanding the context of threat-related terms rather than relying solely on keywords. The system can detect both known and previously unseen threats through unsupervised techniques such as clustering and topic modeling, making it highly adaptive to evolving cyber-attack trends. By integrating data from multiple sources—including blogs, reports, social media, and dark-web platforms—the system generates comprehensive and correlated threat profiles that provide deeper insights into attack patterns. Additionally, the framework is scalable, allowing it to handle continuous data streams and large datasets typical in enterprise environments. Overall, the system improves detection speed, strengthens situational awareness, and supports a more proactive approach to cyber defense.

METHODOLOGY

The proposed system follows a structured NLP-driven approach to automatically identify and profile emerging cyber threats from diverse textual sources. It begins with **data acquisition**, where information is collected continuously from security blogs, threat reports, vulnerability databases, social-media discussions, and dark-web forums. The collected data is then **preprocessed** through tokenization, stop-word removal, lemmatization, and noise reduction to prepare it for analysis. Advanced **NLP techniques**, including transformer-based models such as BERT, are applied for Named Entity Recognition (NER) to extract critical threat indicators such as malware names, CVE identifiers, attack techniques, threat actors, and affected platforms. **Topic modeling** and **semantic similarity analysis** are used to detect emerging patterns and group related indicators. Extracted information is then processed using **machine learning algorithms** to classify threats, cluster related events, and assign severity and novelty scores, enabling the creation of structured threat profiles. Finally, the system generates **analyst-friendly reports** that summarize key indicators, attack trends, severity levels, and potential mitigation measures, while visualization tools help interpret complex relationships. This methodology ensures accurate, real-time, and scalable detection and profiling of emerging cyber threats with minimal human intervention.

System Model

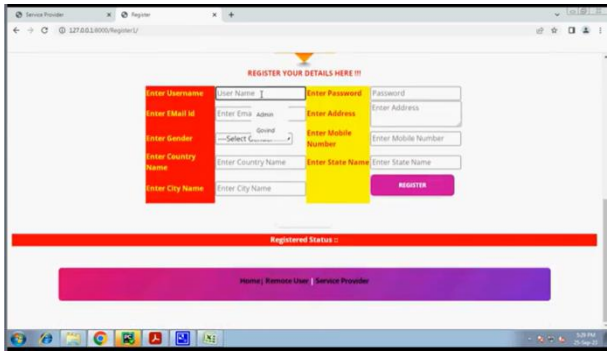
SYSTEM ARCHITECTURE

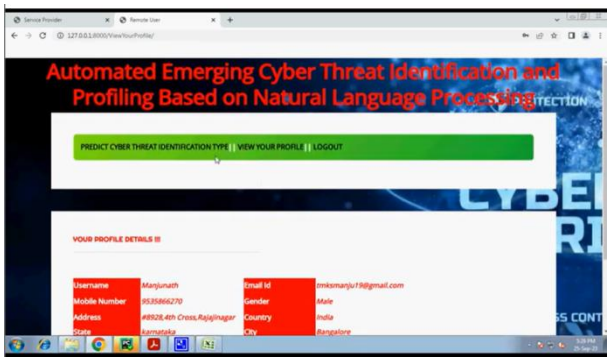


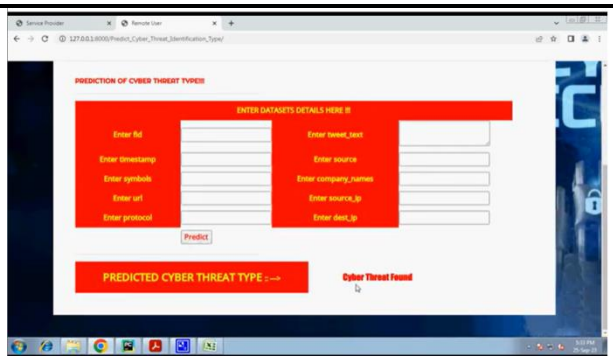
Results and Discussions

Received: 06-03-2026 | Accepted: 14-04-2026 | Published: 21-04-2026

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper







CONCLUSION

The proposed NLP-based framework provides an effective and automated solution for identifying and profiling emerging cyber threats from large volumes of unstructured textual data. Unlike traditional manual or rule-based approaches, the system leverages advanced natural language processing and machine learning techniques to extract relevant threat indicators, detect new attack trends, and generate structured threat intelligence in real time. By integrating multiple data sources—including security blogs, vulnerability databases, social media, and dark-web forums—the framework offers a comprehensive view of potential threats while reducing analyst workload and minimizing detection delays. The use of semantic analysis, topic modeling, and clustering allows the system to recognize previously unseen threats, classify their severity, and provide actionable insights. Overall, this approach enhances situational awareness, strengthens proactive defense strategies, and demonstrates the potential of NLP-driven intelligence systems in modern cybersecurity operations.

REFERENCES

1. Buczak, A. L., & Guven, E. (2016). *A survey of data mining and machine learning methods for cyber security intrusion detection.* IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.
2. Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly detection: A survey.* ACM Computing Surveys, 41(3), 1–58.

Received: 06-03-2026 | Accepted: 14-04-2026 | Published: 21-04-2026

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

3. Sommer, R., & Paxson, V. (2010). *Outside the closed world: On using machine learning for network intrusion detection.* IEEE Symposium on Security and Privacy.
4. Joshi, S., & Thomas, S. (2019). *Natural language processing for threat intelligence: Methods and applications.* Journal of Cybersecurity, 5(1), 1–15.
5. Sadeghzadeh, A., et al. (2021). *AI-driven cybersecurity: Machine learning and NLP applications for threat intelligence.* Journal of Network and Computer Applications, 168, 102739.
6. Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). *Toward developing a systematic approach to generate benchmark datasets for intrusion detection.* Computers & Security, 31(3), 357–374.
7. MITRE Corporation. *ATT&CK Framework.* Retrieved from https://attack.mitre.org/
8. IBM Security. (2020). *The Cost of a Data Breach Report.*
9. Sarker, I. H. (2021). *Machine learning: Algorithms, real-world applications and research directions.* SN Computer Science, 2, 160.
10. Yin, C., et al. (2017). *Deep learning for cyber threat intelligence: Techniques and applications.* Computers & Security, 68, 45–58.