iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** II **Month of publication:** February 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.58497

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

# Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data Using Natural Language Processing

Tamilselvan Arjunan
*Lead Software Engineer*

***Abstract:*** ***Due to the growing volume and variety in data generated by cybersecurity systems, it is crucial that unstructured text be used for detecting anomalies. Natural language processing is a powerful tool for analyzing unstructured information and identifying threats. This paper presents a comprehensive review of NLP applications for cybersecurity. We first present the motivations for and challenges associated with using NLP to improve cybersecurity. Then, we provide background information on unstructured data that is relevant to cybersecurity, and discuss NLP techniques such as named entity recognition (NEAR), sentiment analysis, topic modelling, and document classifying. This paper focuses on how these techniques are used to detect anomalies and intrusions. We present a taxonomy for NLP-driven approaches, and we conduct a literature review that is categorized according to this taxonomy. We examine critically the strengths and weaknesses of current techniques. We highlight research gaps based on this analysis and propose a research agenda to advance NLP research in cybersecurity applications. This paper summarizes previous research and lays the foundation for using NLP to tackle cybersecurity challenges that involve unstructured data.***

***Keywords: Natural Language Processing, Anomaly Detection, Intrusion Detection, and Text Analytics***

## I. INTRODUCTION

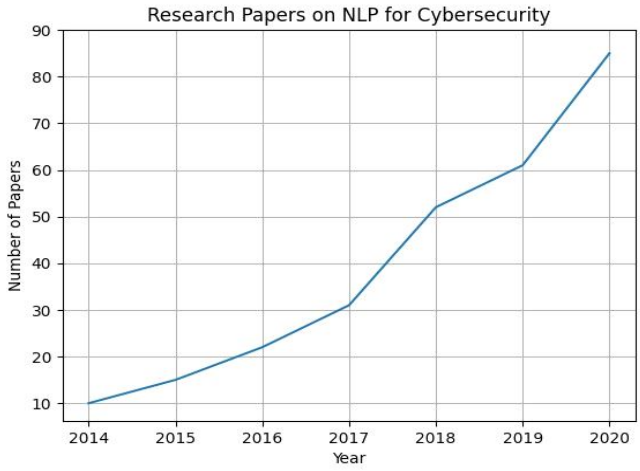
In recent years, the exponential growth of data has led to an era in which unstructured text is a valuable resource for cybersecurity analytics. This vast pool of unstructured information is a collection of data from a variety of sources including, but not limited, to log messages and network traffic texts, social media posts, threats reports, system documentation, etc. Natural Language Processing (NLP), a set of tools designed to extract insights from unstructured text, stands out in this context. NLP's core capabilities are diverse, including named entity recognition and sentiment analysis. Topic modeling, document classification, and topic modeling are also included. The application of NLP in cybersecurity holds promise for facilitating anomaly detection, identifying novel threats, pinpointing misconfigurations, or unauthorized access.

The use of NLP in cybersecurity is not without its challenges. Despite the benefits that NLP can provide, there are several obstacles to overcome. The nuanced meanings and complex semantics of cybersecurity texts are a major obstacle, making it impossible to use standard NLP tools. It is essential to have a deep understanding of domain-specific complexities in order to achieve accuracy for tasks like entity extraction, threat modelling, and intent detection. This problem is compounded by the lack of cybersecurity corpora that are essential to train robust NLP models tailored for the intricacies within the cybersecurity domain. NLP persists despite these obstacles, due to the wealth and depth of insights contained within unstructured data.

This paper's primary goal is to provide a comprehensive examination of the use of NLP in the context of cybersecurity applications. The paper aims to contribute in several ways.

*1)* Examine the reasons for the adoption of NLP within cybersecurity, while also highlighting the challenges that have been encountered.
*2)* Provide an in-depth analysis of the various unstructured data resources relevant to cybersecurity. This will provide a solid understanding of the data environment.
*3)* Give a comprehensive review of the NLP methods that are relevant to cybersecurity applications. Describe their functionality and possible applications.
*4)* Create a systematic taxonomy describing NLP-driven anomaly detection and intrusion detection methods, providing a structured framework to understand and categorize existing research efforts.

©IJRASET: All Rights are Reserved 




This paper is a synopsis of existing literature and explores relevant technical frameworks and avenues for further exploration. It aims to be a catalyst to advance research and encourage widespread adoption NLP techniques in order to tackle the cybersecurity challenges that are associated with unstructured information [5].

## II. UNSTRUCTURED DATA SOURCES

This article provides an overview of the key sources of unstructured data that are relevant to cybersecurity applications.

System logs: A veritable treasure trove for security insights:System Logs are the unsung heroes in the security landscape. They provide invaluable insight into the inner workings and applications of operating systems. Although often ignored, these voluminous records contain a wealth information in semi-structured form. Imagine a tapestry of structured fields like timestamps or event IDs interwoven with rich narratives of free-form unstructured text messages. Textual components are the lifeblood of logs and can be used to identify anomalies or potential security threats.

When you dig deeper, you'll find that the text messages in system logs usually include a wide range of entries. They provide a comprehensive picture of system activities, from routine operational information to critical error messages. The meticulously recorded technical details provide valuable insight into the system configuration, software version, and resource usage. Security professionals are alerted to possible issues by warnings that act as sentinels. These textual elements, when transformed into actionable intelligence using NLP, can help defenders detect malicious activity before it becomes a full-blown incident. To unlock the full potential of system logs, you will need to overcome certain obstacles. Semi-structured data requires specialized parsing methods that can handle both structured fields as well as the subtleties of human language. The sheer volume of data generated by logs can be overwhelming and require efficient methods for filtering and analyzing. The insights gained from logs are often game-changing and can empower security.

Teams can proactively identify threats and take action to ensure the smooth operation of their systems and a robust security posture.

Network traffic: Decoding the whispers on the wire:Network Traffic, the lifeblood for the digital world, pulses with information. Beyond the raw data flowing through these channels, there is another layer of intelligence: the unstructured texts embedded in network packets. These packets are digital envelopes that carry data. They have headers and payloads which contain valuable secrets.

Table 1: Comparison of Traditional vs. NLP-based Techniques for Intrusion Detection

| Feature | Traditional Techniques | NLP-based Techniques |
|---|---|---|
| Data Type | Structured (network logs, system data) | Unstructured (textual data) |
| Attack Detection Scope | Signature-based, known attacks | Behavioral analysis, zero-day attacks |
| Adaptability | Limited, require rule updates | Continuously learns and adapts to new patterns |
| False Positives | High due to rigid rules | Potentially lower due to context understanding |

But extracting useful insights from the text of network traffic presents its own challenges. Real-time analytics are required to keep up with the constantly changing data flow. The sheer volume of data can also be overwhelming. This requires efficient filtering and priority techniques to focus only on the most important information. The ability to gain insights from the network traffic text allows security teams to respond in real time to threats, protecting the integrity and confidentiality data traveling the digital highways.

Threat Reports: In an ever-evolving game of cybersecurity chess, threat reports are invaluable, providing a peek into the playbook of the adversary. These reports, carefully crafted by cybersecurity agencies and firms, serve as sentinels to alert defenders about the latest tools, techniques, and vulnerabilities used by malicious actors. The true power of these reports is not only in the structured data that they provide, but in the rich textual descriptions that are woven throughout their pages.

The narratives are authored by humans and paint a vivid image of hacker operations. They detail their tactics, procedures, and techniques (TTPs). These descriptions are often filled with technical jargon, domain-specific terms, and other specialized terminology. They provide a crucial context to understand the nature of the attack. Security professionals can extract key indicators (IOCs), such as file names, registry entries, or network commands, from these textual descriptions by leveraging NLP.

Social Media: While social media's surface is filled with harmless updates and playful interaction, the dark corners of hacker platforms and forums are a thriving underbelly. Text is freely exchanged in these secretive digital spaces. Discussions on attacks, vulnerabilities, tools and techniques are discussed with alarming candor. This seemingly harmless chatter can be a valuable resource for situational awareness, and an early warning system against emerging threats. Imagine a bustling market of illicit information, where rivals exchange whispered secret messages in text form. The details of the trades, debates, tools, and attack methods are all meticulously documented. Security professionals can access this secret world by harnessing NLP. They will gain insights into the most recent threats and anticipate the tactics of their adversaries. The textual discussion within these forums may reveal valuable information.

Emerging vulnerabilities. Hackers discuss newly discovered security vulnerabilities before they become public, giving defenders a critical window of time to patch their systems.

Security professionals can proactively protect themselves against attacks by reading detailed descriptions of the attack methods and procedures.

Discussions about threat actors allow for identification of active groups, their areas and focus, as well as potential targets.

Navigating the murky waters on social media to achieve cybersecurity goals presents a unique set of challenges. To identify the relevant conversations amongst all of the noise, filtering and analytic techniques are needed to deal with such a large volume of data. The dynamic nature of these platforms also requires real-time monitoring to stay on top of emerging threats. In addition, to operate within these spaces legally and ethically requires careful consideration as well as adherence with platform regulations. The insights gained from text data on social media can be a valuable asset in the ongoing fight against cybercrime.

Cybersecurity Corpora : In cybersecurity, structured information is king. It's meticulously organized into standardized formats such as STIX (Structured Threat Information eXchange), MAEC (Malware Attribute Enumeration, Characterization and Classification) for attack patterns and OASIS Open Threat Modeling (Open Threat Model) to model threats. These frameworks are a great way to understand and mitigate cyber threats. Within this structured world, there is another layer of intelligence: the unstructured text embedded within.

Imagine structured knowledge is the skeleton or framework of a structure, which provides essential support and organisation. Unstructured text is the blood and flesh that gives the structure life. Textual narratives in threat intelligence reports provide rich context and human insight that can't be captured by structured fields. Textual explanations are often included with MAEC attack pattern descriptions to illuminate the attackers' intent and capabilities. Open Threat Model narratives also provide context to help understand the attack vectors that could be associated with a particular system.

Security professionals can unlock hidden potential in embedded text by harnessing NLP. The narratives in STIX reports may contain indicators of compromise that would be missed by structured analysis. MAEC attack patterns can be used to extract insights that help understand attacker motivations and behavior. Textual analysis in Open Threat Models is also a powerful tool for identifying potential vulnerabilities and guiding mitigation strategies. Working with embedded text in cybersecurity corpora poses its own challenges. NLP models need to be trained in specialized terminology. Interoperable solutions are also required to bridge the gap that exists between text and structured data. The ability to harness embedded text in cybersecurity corpora can empower defenders to gain a more nuanced and richer understanding of threats, leading to a more effective security posture [17].

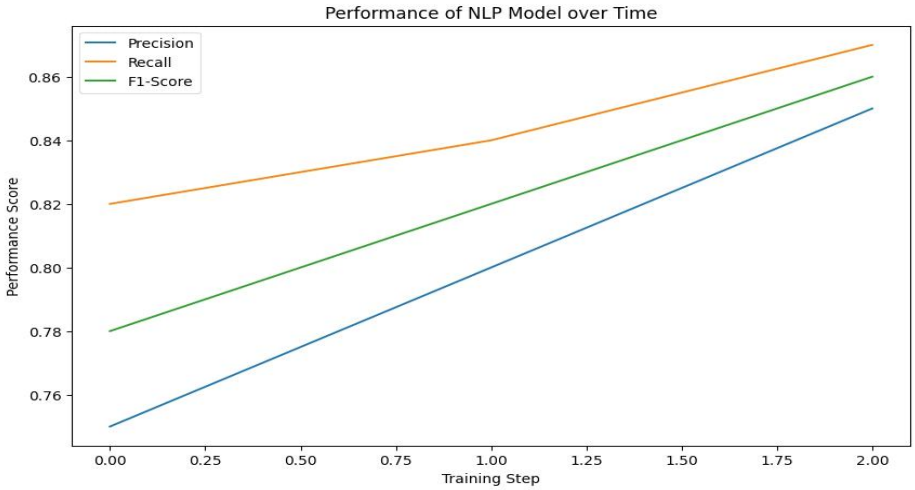
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

## III. TAXONOMY OF NLP APPLICATION AREAS

We present a taxonomy for NLP techniques used in anomaly detection and intrusion prevention - both core cybersecurity capabilities.

Anomaly Detection Anomaly Detection identifies events or activities that are unusual and deviate away from normal patterns. NLP improves anomaly detection by the following methods:

Log Detection of Anomalies: Log anomaly detection is a key component of Natural Language Processing in system monitoring. It involves several key methods. The extraction of log templates using clustering techniques allows the creation of profiles of normal log patterns. These templates are used to identify deviations from expected behavior. The generated log templates can also be used to identify outliers in the log data. A second method involves embedding the log messages in a vector space and then applying reconstruction techniques to identify anomalies on the basis of deviations from the reconstructed images. Log messages can also be classified in order to identify abnormal sequences, missing events and unknown message types. This multifaceted approach for log anomaly detection uses advanced NLP techniques in order to monitor and identify irregularities with system log data. It enhances the overall security of the system.




Network Analysis of Traffic: Log anomaly detection is an important aspect of network security. This is especially true in the field of Natural Language Processing. A common approach is to analyze the text in network packets for anomalies. This can be achieved by using various techniques such as extracting named entities and protocol grammar from the packet payload. As outlined in the reference, this allows for detection of protocol violations and misconfigurations. As highlighted in the reference [18], another method involves classifying web traffic on HTTP requests and distinguishing between malicious domains. As discussed in the reference, SMTP sender/recipient and email content analysis can also be used to detect spam or phishing. A second technique that is worth mentioning involves the training of sequence models using normal traffic patterns, and then using reconstruction error in order to identify anomalies. These methods collectively enhance network security by proactively identifying potential threats and vulnerabilities.

Threat Analysis of Intelligence: In the world of cybersecurity, log anomaly detection plays a crucial role in identifying new threats and vulnerabilities. Natural Language Processing (NLP) is one method used to identify anomalies in cyber threat reports, disclosures on hacker forums and other sources. This is done by using Named Entity Recognition and relation extraction techniques in order to identify new threat actors, campaigns, and tactics, techniques, and procedures (TTPs). Topic-based analysis is also used.

Modeling allows for the detection of changes in hacker discussion, which may indicate the emergence or new threats. Analyzing the timelines of threats reports helps identify sudden spikes in vulnerabilities and exploits. This allows for timely responses to possible threats. Hacker sentiments are classified as either positive, neutral or negative. This helps to prioritize the severity of vulnerabilities. This comprehensive methodology combines various NLP techniques in order to improve the detection and priority of cyber threats within an ever-evolving environment [20].

Intrusion Detection:Intrusion-detection systems (IDSs) detect unauthorized access, policy breaches, and attacks. NLP enhances IDS capabilities by:

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Attack Detection of Patterns: The process involves translating unstructured information into structured attack patterns, indicators of compromise and thereby making it easier to identify known threats. This can be achieved in several ways:

First, extracted entities, such as software and vulnerabilities, are linked with attack templates and threat dictionary. This linkage allows the recognition of known attacks patterns and indicators that are associated with these entities.

Second, documents are classified into attack taxonomy classes using multi-label classification techniques. This categorization helps to understand the content of the document and its relevance for specific types of attacks.

Finaly, the network traffic text is converted into vectors and distance similarity matching methods are used to identify patterns that indicate known malicious activities. Comparing the characteristics of network activity to known bad traffic signatures can flag suspicious activities for further investigation.

Malicious Detection of Content: In the domain for malicious content detection, various techniques are used to identify threats, such as malware and phishing attempts. The system relies primarily on textual signals. The system uses lexical features to classify domains, URLs and network traffic into malicious or benign. The system can also detect obfuscated C2 requests and domain DNS queries that may indicate the presence of Domain Generation Algorithm (DGA) algorithms. It also identifies phishing by performing natural language processing on URLs and email body content.

These patterns and characteristics are indicative of phishing. These methods together contribute to a robust defence mechanism against malicious content of various types, improving overall security posture.

Insider Detection of Threats: One crucial aspect is monitoring internal communications and logs in order to detect potential threats from rogue employees. As indicated by research [22], this requires a multi-faceted approach that includes the classification of employee communications and logins in order to identify disgruntled employees. As highlighted in a second study [23], the analysis of logins over varying timeframes and geographic locations can also help detect anomalous activity. As discussed in research [24], linking extracted username entities with HR databases is a useful strategy for identifying unauthorized use of credentials. This taxonomy provides a systematic framework to understand the primary applications for Natural Language Processing (NLP). We will now consolidate the key insights from existing literature on these specific topics [23].

## IV. RESEARCH GAPS AND OUTLOOK

The research in NLP for cybersecurity is facing several challenges and opportunities for advancements. The creation of representative datasets that span diverse cybersecurity textual resources is a crucial aspect. These datasets are crucial for a robust evaluation and comparison NLP techniques [29]. The challenge is curating datasets to include the variety of language styles and contexts in cybersecurity texts.

Table 3: Real-world Case Studies of NLP-based Intrusion Detection

| Case Study Social | Data Source | Threat Detected Malicious URLs | Outcome |
|---|---|---|---|
| Media Phishing | Tweets, forum posts links | disguised as legitimate | Timely intervention, prevented user clicks |
| Email Threat Detection | Phishing emails, spear phishing attempts | Suspicious language patterns and urgency tactics | Early detection, avoided data breaches |
| Insider Threat | Internal communication | Unusual collaboration between accounts, abnormal | Identified potential insider activity, keyword enabled |

The development of flexible, unsupervised or semi-supervised methods is another key research area. These methods must be able to adapt to new threats effectively without having to rely on large labeled sets of training, which are not always available or easy to obtain. This adaptability is essential in the rapidly changing cyber threat landscape where new attack patterns and vectors are constantly emerging [30]. The advancement of graph-based methods will also be crucial for capturing complex relationships between cyber entities described in textual data. The graph-based representations provide a powerful framework to model interconnected entities, their interactions and a holistic view on cyber threats. Further research is required to improve the scalability of these methods for large-scale applications in cybersecurity.

NLP can be used to explore sources that are not being utilized, such as hacker forums or threat reports, for the purpose of identifying leading indicators. These sources can provide valuable information and early warnings on emerging threats and vulnerabilities. Using NLP techniques to extract information from these sources is a valuable way to provide intelligence for efforts in threat detection and mitigation. It is important to enhance entity linking and knowledge-graph techniques that are tailored for cybersecurity ontology. Effective entity linking allows for the identification and resolutions of entities (such organizations, malware or attack techniques), mentioned in cybersecurity texts. This facilitates deeper analysis and understanding the threat landscape. Knowledge graph techniques can also be used to organize and represent cybersecurity information in a structured, interconnected way, which allows for more effective reasoning.

Another research priority is to design end-to-end NLP systems that combine multiple techniques in order to improve detection accuracy. To achieve robust and reliable detection, such pipelines should include various stages of text-processing including preprocessing. Feature extraction, modeling and postprocessing.

Cybersecurity threats. It is important to develop NLP models which are easily understood by cybersecurity experts in order to foster trust and the adoption of NLP technologies for security applications. These models must provide transparency into the decision-making processes, which will allow analysts to interpret their findings and validate them effectively. Another important research area is to enable active learning by incorporating human feedback. Active learning techniques enable NLP models iteratively to improve their performance based on feedback from humans. This method can reduce the burden of annotation and speed up the development of NLP models.

It is crucial to assess the resilience of NLP based cybersecurity systems in the face of adversarial attacks by creating realistic adversarial samples generation methods. Adversarial inputs are carefully crafted to fool NLP models. It is crucial to develop realistic adversarial sample that mimics real-world threats in order to accurately evaluate the robustness NLP models.

Finaly, it is important to accelerate research and development by facilitating adoption of open-source libraries and guidelines for NLP in security. Open-source resources offer a framework that allows researchers and practitioners alike to collaborate, exchange insights and build on each other's works, ultimately driving progress and innovation in cybersecurity NLP.

## V. CONCLUSION

This paper examines the use of natural language processing in the domains of anomaly and intrusion detection. We have explored the reasons behind the adoption of NLP within such a complex environment, and the challenges associated with its application. We began our examination with an assessment of the different sources of unstructured data that are relevant to cybersecurity monitoring. This set the stage for a more in-depth exploration of NLP.

We explained fundamental NLP techniques, including named entity recognition and sentiment analysis. Topic modeling and document classification were also discussed. These techniques could be used to enhance anomaly and intrusion detector efforts [32]. We aimed to provide a taxonomy that outlines the various ways NLP capabilities can enhance detection mechanisms.

Comprehensive understanding of their impact. This paper also conducted an extensive literature review, synthesizing the key insights gained from applying NLP techniques to diverse data sources, including system logs and network traffic. This synthesis was able to not only demonstrate the effectiveness of NLP as a security measure but also highlight research gaps, and outline a roadmap for further investigations.

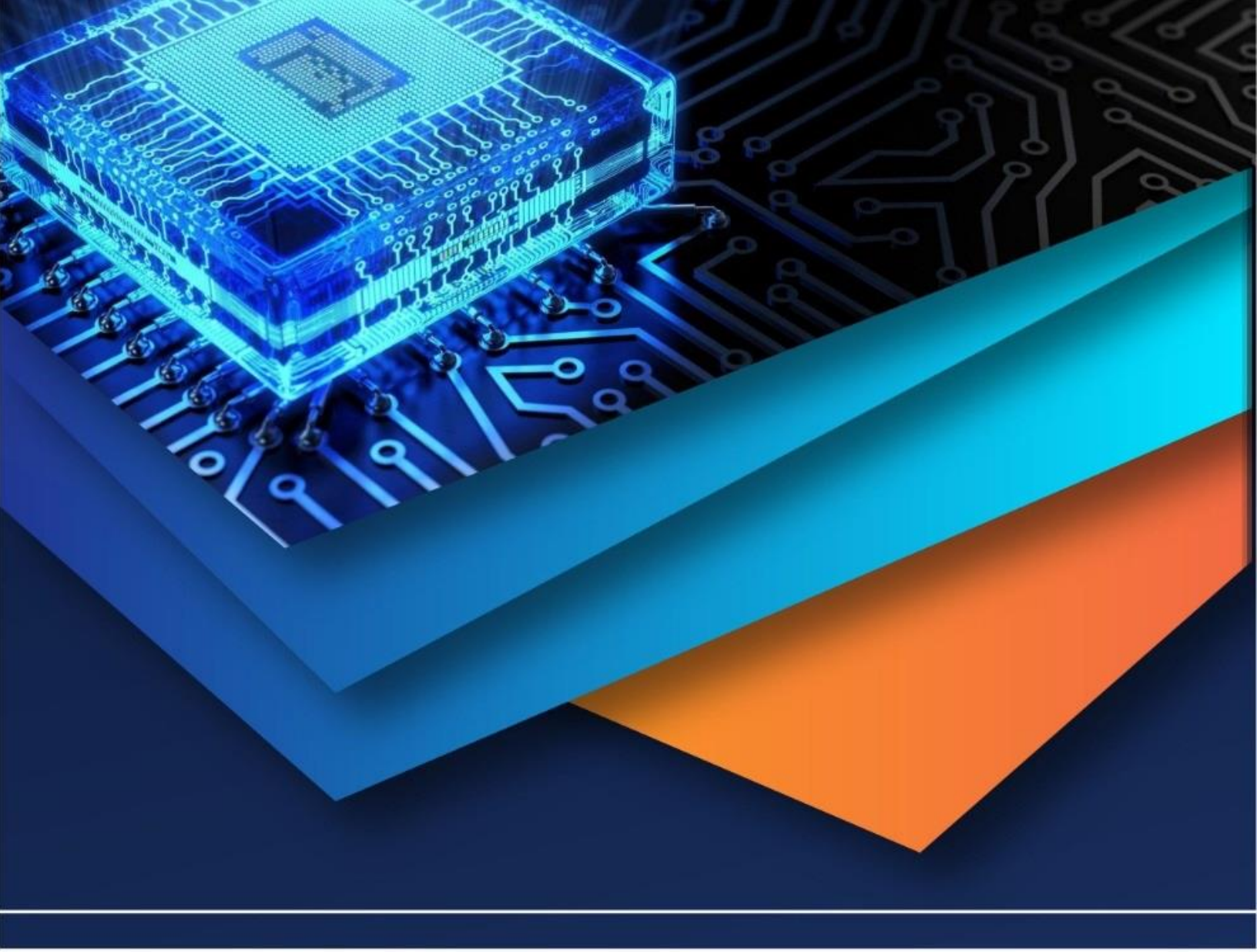
The importance of NLP for extracting actionable intelligence out of unstructured data is not to be underestimated. Cyber threats are becoming more sophisticated and diverse. Its ability for cybersecurity practitioners to extract meaningful insights and discern patterns makes it an indispensable tool. Utilizing NLP's full potential can help organizations strengthen their defenses by enabling them to detect, investigate, and respond to emerging threats.

The future of NLP for cybersecurity is bright, as research and innovation efforts are poised to unlock greater efficiencies and abilities. The cybersecurity community can use NLP's full potential to protect digital assets and maintain the integrity of critical infrastructures by addressing identified research challenges, and exploring avenues for improvement and advancement.

## REFERENCES

[1] X. Wang, Z. Xu, and X. Gou, "The Interval probabilistic double hierarchy linguistic EDAS method based on natural language processing basic techniques and its application to hotel online reviews," Int. J. Mach. Learn. Cybern., vol. 13, no. 6, pp. 1517–1534, Jun. 2022.
[2] A. Lavecchia, "Deep learning in drug discovery: opportunities, challenges and future prospects," Drug Discov. Today, vol. 24, no. 10, pp. 2017–2032, Oct. 2019.
[3] T. K. Mackey et al., "Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram," JMIR Public Health Surveill., vol. 6, no. 3, p. e20794, Aug. 2020.
[4] I. Doghudje and O. Akande, "Dual User Profiles: A Secure and Streamlined MDM Solution for the Modern Corporate Workforce," JICET, vol. 8, no. 4, pp. 15–26, Nov. 2023.
[5] K. N. Syeda, S. N. Shirazi, S. A. A. Naqvi, H. J. Parkinson, and G. Bamford, "Big Data and Natural Language Processing for analysing railway safety," in Innovative Applications of Big Data in the Railway Industry, IGI Global, 2018, pp. 240–267.
[6] S. Thejaswini and C. Indupriya, "Big data security issues and natural language processing," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019.
[7] M. Khader, A. Awajan, and G. Al-Naymat, "The effects of natural language processing on big data analysis: Sentiment analysis case study," in 2018 International Arab Conference on Information Technology (ACIT), Werdanye, Lebanon, 2018.
[8] J. P. Singh, "Mitigating Challenges in Cloud Anomaly Detection Using an Integrated Deep Neural Network-SVM Classifier Model," Sage Science Review of Applied Machine Learning, vol. 5, no. 1, pp. 39–49, 2022.
[9] H. M. Khan, F. M. Khan, A. Khan, M. Z. Asghar, and D. M. Alghazzawi, "Anomalous Behavior Detection Framework Using HTM-Based Semantic Folding Technique," Comput. Math. Methods Med., vol. 2021, p. 5585238, Mar. 2021.
[10] D. R. Harris, C. Eisinger, Y. Wang, and C. Delcher, "Challenges and barriers in applying natural language processing to medical examiner notes from fatal opioid poisoning cases," Proc. IEEE Int. Conf. Big Data, vol. 2020, pp. 3727–3736, Dec. 2020.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)