Int. J. Inf. Cybersec.-2023

Natural Language Processing for Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data

Suresh Sharma

Department of Cybersecurity Analytics, Tribhuvan University, Nepal

suresh.sharma@tribhuvanuniversity.edu.np

Tamilselvan Arjunan

arjunantamilselvan1@gmail.com

Abstract

With the increasing volume and variety of data generated in cybersecurity systems, leveraging unstructured text data has become crucial for detecting anomalies and intrusions. Natural language processing (NLP) provides effective techniques for analyzing unstructured data and identifying threats. This paper provides a comprehensive overview of NLP techniques for cybersecurity applications. First, we present the motivations and challenges of using NLP in cybersecurity. We then provide background on the types of unstructured data relevant to cybersecurity and discuss NLP methods including named entity recognition, sentiment analysis, topic modeling, and document classification. The core of the paper examines how these techniques can be used for anomaly detection and intrusion detection systems. We provide a taxonomy of NLP-driven approaches and conduct an extensive literature review categorized along this taxonomy. We critically examine the advantages and limitations of current techniques. Based on this analysis, we highlight research gaps and propose an agenda for advancing NLP research for cybersecurity applications. Overall, this paper synthesizes past research and establishes a foundation for applying NLP to address pressing cybersecurity challenges involving unstructured data.

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

Introduction

The exponential proliferation of data in recent years has ushered in an era where unstructured text data emerges as an invaluable resource for cybersecurity analytics. This expansive pool of unstructured data encompasses a myriad of sources, including but not limited to log messages, network traffic text, social media posts, threat reports, and system documentation [1]. Within this landscape, Natural Language Processing (NLP) stands out as a critical set of techniques designed to unlock insights from such unstructured text data. The core capabilities of NLP span a wide spectrum, encompassing named entity recognition, sentiment analysis, topic modeling, and document classification, among others. In the realm of cybersecurity, the application of NLP holds promise in facilitating the detection of anomalies, uncovering novel threats, and pinpointing misconfigurations or unauthorized access [2].

However, despite the potential benefits, several formidable challenges loom on the horizon, hindering the effective utilization of NLP for cybersecurity endeavors. One of the foremost obstacles lies in the intricate semantics and nuanced meanings pervasive within cybersecurity text, rendering the straightforward application of off-the-shelf NLP tools insufficient. A deep understanding of the domain-specific intricacies is imperative to ensure the accuracy of tasks such as entity extraction, threat modeling, and intent detection. Compounding this issue is the scarcity of labeled cybersecurity corpora, essential for training robust NLP models tailored to the intricacies of the cybersecurity domain [3], [4]. Yet, notwithstanding these impediments, NLP perseveres as a promising avenue, owing to the wealth of invaluable insights harbored within unstructured data.

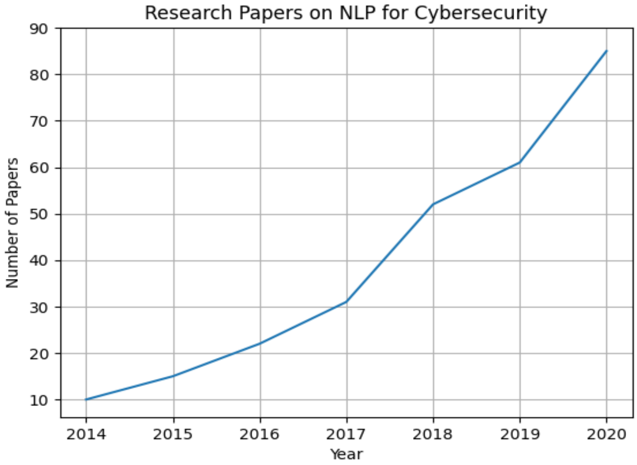
The primary objective of this paper is to furnish a comprehensive exploration of the utilization of NLP techniques within the realm of cybersecurity applications. To this end, the paper endeavors to make several significant contributions:

1) Delve into the motivations driving the adoption of NLP in cybersecurity while elucidating the challenges encountered along the way.

2) Offer an in-depth examination of the diverse array of unstructured data sources relevant to cybersecurity, thereby providing a foundational understanding of the data landscape.

3) Provide a comprehensive overview of the key NLP methodologies pertinent to cybersecurity applications, delineating their functionalities and potential applications.

4) Construct a systematic taxonomy outlining NLP-driven anomaly and intrusion detection approaches, thereby offering a structured framework for understanding and categorizing existing research endeavors.



By synthesizing existing literature, elucidating pertinent technical frameworks, and delineating avenues for future exploration, this paper aspires to serve as a catalyst for advancing research efforts and fostering widespread adoption of NLP techniques to address the pressing cybersecurity challenges associated with unstructured data [5].

Motivations and Challenges

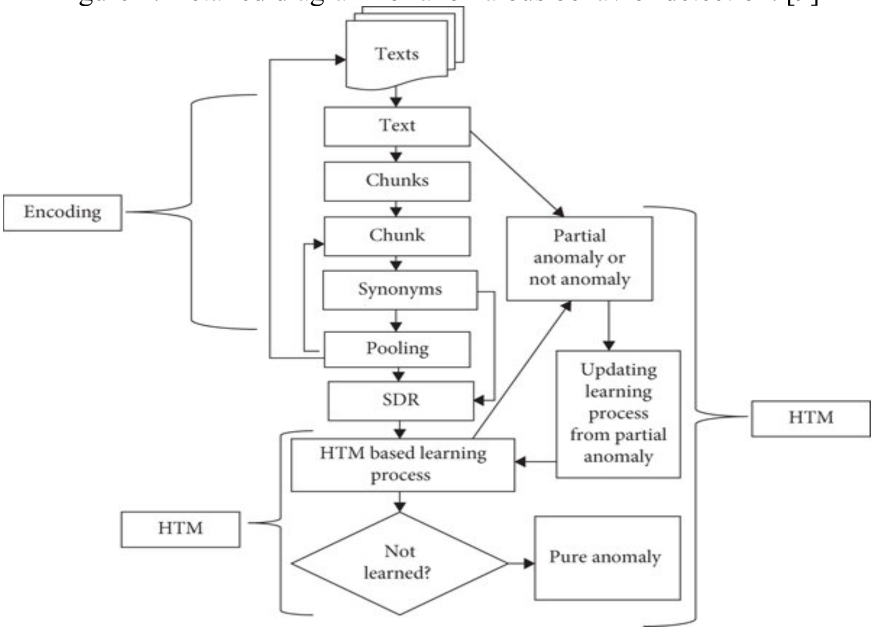
In this section, we discuss the key motivations for using NLP in cybersecurity and the main challenges faced.

Motivations: The burgeoning field of cybersecurity faces the daunting challenge of safeguarding an ever-expanding digital landscape. Traditional approaches, while providing a vital foundation, often struggle to keep pace with the sophistication and dynamism of modern cyber threats [6]. In this context, Natural Language Processing (NLP) emerges as a powerful tool, offering compelling motivations for its adoption as a potent weapon in the cybersecurity arsenal.

Firstly, NLP possesses the unique ability to unlock the vast potential of unstructured data, which constitutes a staggering 80-90% of the information within the cybersecurity domain [1]. This data trove, encompassing threat reports, social media chatter, email communications, and other textual resources, remains largely untapped due to its inherent lack of structure [7]. NLP techniques, however, empower analysts to extract valuable insights from this unstructured data, unveiling hidden patterns, uncovering emerging threats, and gleaning critical details often overlooked by traditional analysis methods [8].

Furthermore, NLP plays a pivotal role in early anomaly detection. Traditional signature-based methods often lag behind new and evolving threats, leaving systems vulnerable in the critical window before detection. NLP, on the other hand, analyzes textual data for subtle anomalies, linguistic deviations, and emerging threat indicators, enabling proactive detection before these threats manifest in structured data like network logs or file access attempts. This increased lead time empowers defenders to mitigate threats swiftly and effectively, minimizing potential damage.

Figure 1. Detailed diagram for anomalous behavior detection. [9]

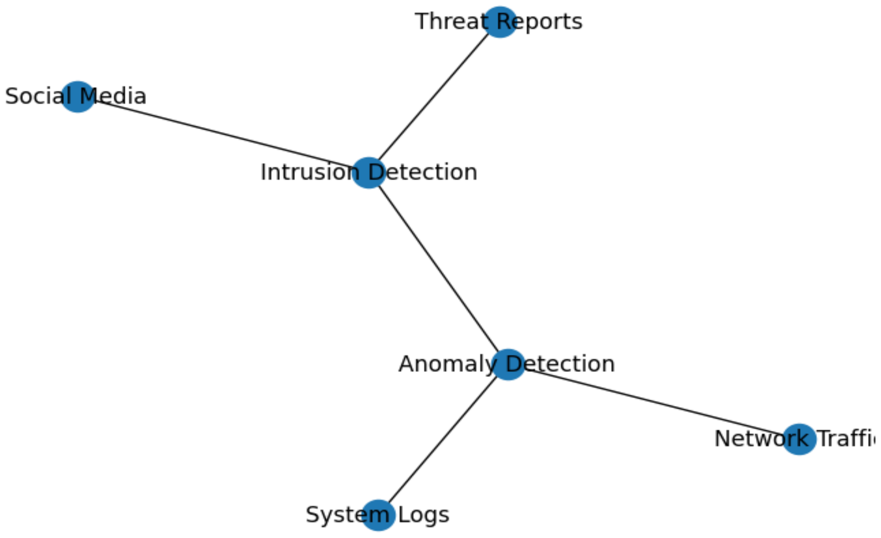

Texts
Text
Chunks
Chunk
Synonyms
Pooling
SDR
HTM based learning process
Not learned?
Pure anomaly
Partial anomaly or not anomaly
Updating learning process from partial anomaly
Encoding
HTM
HTM

Beyond early detection, NLP fosters a holistic situational awareness within the security landscape. By analyzing threat reports, news articles, and other text-based sources, NLP paints a comprehensive picture of the evolving threat environment. This broader perspective allows defenders to anticipate trends, identify emerging adversaries, and understand the motivations and tactics employed by malicious actors. Such comprehensive understanding proves invaluable in crafting effective mitigation strategies and prioritizing security efforts [10]. Moreover, NLP transcends the limitations of traditional rule-based and signature-based detection methods. These methods struggle to identify novel threats that deviate from predefined patterns. NLP, in contrast, leverages its linguistic prowess to detect emerging threats even if they exhibit novel characteristics or employ previously unseen tactics. This adaptability ensures that security measures remain agile and responsive in the face of a constantly evolving threat landscape [11].

Finally, NLP offers significant efficiency benefits. Manual analysis of vast textual datasets is both time-consuming and resource-intensive. NLP automation, however, streamlines the process, enabling defenders to analyze large volumes of text data swiftly and efficiently. This newfound efficiency allows for continuous monitoring and analysis, ensuring that critical information is not overlooked and timely insights are readily available to inform security decisions [12].

Challenges: While the allure of NLP's potential in cybersecurity is undeniable, its practical implementation presents a formidable array of challenges that demand innovative solutions.

Decoding Semantic Intricacies: The very nature of cybersecurity discourse presents the first hurdle. Technical jargon, domain-specific acronyms, and intricate sentence structures abound, often posing significant parsing challenges for standard NLP models. Understanding the nuanced interplay of these linguistic elements is crucial for accurate information extraction and threat detection. Understanding Specialized Entities: Cybersecurity threats lurk within a unique ecosystem of entities - exploits, malware families, vulnerabilities, and protocols. Distinguishing these entities from generic terms and accurately capturing their relationships within the text is essential for comprehending the true nature of potential threats. This necessitates the development of specialized NLP models trained on domain-specific corpora and equipped with entity recognition capabilities.



Data Scarcity, A Training Impediment: The supervised learning paradigm, prevalent in many NLP applications, hinges on the availability of large, labeled datasets for model training. Unfortunately, such labeled corpora are scarce in the cybersecurity domain, hindering the development of robust and generalizable NLP models [13]. This necessitates exploring alternative learning paradigms, such as transfer learning and semi-supervised learning, to leverage limited labeled data effectively. Adversarial Evasion: A Deceptive Dance: Cyber adversaries, ever-evolving in their ingenuity, actively craft attacks designed to circumvent NLP detection. Obfuscation techniques, synonymous substitution, and syntactic manipulation pose significant challenges, demanding the development of robust NLP models capable of identifying and neutralizing these deceptive tactics [14].

Accuracy versus Interpretability: A Delicate Balance: While black-box NLP models often achieve impressive accuracy, their lack of interpretability renders them unsuitable for security applications. Understanding the rationale behind NLP-driven decisions is crucial for building trust and enabling human oversight in critical security scenarios. Striking a balance between accuracy and interpretability remains a key research area. Evaluation Conundrum: Gauging Operational Fitness: Assessing the effectiveness of NLP-based solutions in a real-world cybersecurity context presents a unique challenge. Traditional evaluation metrics might not adequately capture the nuances of operational deployment. Developing robust and domain-specific evaluation methodologies is essential for ensuring the practical efficacy of NLP solutions [15].

These challenges, though formidable, serve as a springboard for ongoing research efforts. By tailoring NLP methodologies to the intricacies of the cybersecurity domain, researchers are striving to unlock the full potential of NLP in safeguarding our digital landscape. The next section delves into the diverse and valuable sources of textual data that fuel these advancements.

Unstructured Data Sources

We provide an overview of key sources of unstructured text data relevant for cybersecurity applications.

System Logs: A Veritable Treasure Trove for Security Insights: System logs, the unsung heroes of the security landscape, are invaluable chonicles of the inner workings of operating systems, applications, and network devices. These voluminous records, though often overlooked, harbor a wealth of information in the form of semi-structured data. Imagine a tapestry woven with threads of structured fields, like timestamps and event IDs, intertwined with the rich narrative of unstructured free-form text messages. This textual component, the very lifeblood of system logs, often holds the key to uncovering anomalies and thwarting potential security threats.

Delving deeper, the text messages within system logs typically encompass a diverse range of entries. From routine operational details to critical error messages, they paint a detailed picture of the system's activities. Technical details, meticulously recorded, provide valuable insights into system configurations, software versions, and resource utilization. Warnings, like sentinels standing guard, alert security professionals to potential issues, such as failed login attempts or suspicious file access patterns. By harnessing the power of NLP, these textual elements can be transformed from raw data into actionable intelligence, enabling defenders to detect malicious activity in its early stages before it escalates into a full-blown security incident. However, unlocking the true potential of system logs requires navigating certain challenges. The semi-structured nature of the data necessitates specialized parsing techniques capable of handling both structured fields and the nuances of human language. Additionally, the sheer volume of log data can be overwhelming, demanding efficient filtering and analysis methods. Despite these hurdles, the insights gleaned from system logs can be game-changing, empowering security teams to proactively identify and address threats, ensuring the smooth operation and robust security posture of their systems.

Network Traffic: Decoding the Whispers on the Wire: Network traffic, the lifeblood of the digital world, pulsates with information flowing through the veins of our interconnected systems. But beyond the raw data coursing through these channels lies another layer of intelligence - the unstructured text embedded within network packets. These packets, the digital envelopes carrying data, possess headers and payloads that whisper valuable secrets in the form of text.

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
Interpretability	Black box approach	Explainable reasoning behind detections

Unstructured text headers, laden with protocol commands and error messages, offer glimpses into the technical conversations taking place between systems. Analyzing these headers can reveal anomalous network activity, such as unauthorized connections or attempts to exploit vulnerabilities. Delving deeper, the payloads of certain plaintext protocols, like SMTP and HTTP, unveil a treasure trove of human-readable data. Emails, web requests, and other communication channels flow through these protocols, leaving behind textual traces that can be deciphered with the aid of NLP. By dissecting these textual elements, security professionals can uncover malicious activity hidden within seemingly innocuous exchanges, such as phishing attempts embedded in emails or unauthorized data transfers concealed within web requests. However, extracting meaningful insights from network traffic text presents its own set of challenges. The dynamic nature of network communication necessitates real-time analysis capabilities to keep pace with the ever-evolving flow of data. Additionally, the sheer volume of traffic can be overwhelming, demanding efficient filtering and prioritization techniques to focus on the most critical information. Despite these challenges, the ability to glean insights from network traffic text empowers security teams to detect and respond to threats in real-time, safeguarding the integrity and confidentiality of data traversing the digital highways [16].

Threat Reports: In the ever-evolving chess game of cybersecurity, threat reports serve as invaluable intel, offering a glimpse into the adversary's playbook. These detailed reports, meticulously crafted by cybersecurity firms and agencies, act as sentinels, warning defenders of the latest vulnerabilities, tools, and techniques employed by malicious actors. But the true power of these reports lies not just in the structured data they present, but also in the rich tapestry of textual descriptions woven within their pages.

Extensive narratives, penned by human analysts, paint a vivid picture of hacker operations, detailing their tactics, techniques, and procedures (TTPs). These descriptions, often laden with technical jargon and domain-specific terminology, offer crucial context for understanding the nature of the threat. By leveraging NLP, security professionals can extract key indicators of compromise (IOCs) from these textual descriptions, such as specific file names, registry keys, or network commands associated with malicious activity.

Social Media: While the surface of social media glitters with innocuous updates and playful interactions, a hidden underbelly thrives in the shadowy corners of hacker forums and platforms. Within these clandestine digital spaces, text flows freely, carrying discussions on vulnerabilities, attacks, tools, and techniques with an alarming candor. For the discerning eye, however, this seemingly innocuous chatter transforms into a valuable source of situational awareness and early warnings against emerging threats. Imagine a bustling marketplace of illicit knowledge, where adversaries exchange whispered secrets in the form of text. Exploits are traded, attack methods debated, and tools bartered, all documented in intricate detail. By harnessing the power of NLP, security professionals can gain unprecedented access to this clandestine world, gleaning insights into the latest threats and anticipating the evolving tactics of adversaries. Analyzing textual discussions within these forums can reveal valuable information about:

Emerging vulnerabilities: Hackers often discuss newly discovered vulnerabilities before they are publicly known, providing a crucial window of opportunity for defenders to patch their systems.

Attack methods: Detailed descriptions of attack techniques and procedures (TTPs) shared within these forums equip security professionals with the knowledge to proactively defend against them.

Threat actor activity: Monitoring discussions allows for the identification of active threat groups, their areas of focus, and potential targets.

However, navigating the murky waters of social media for cybersecurity purposes presents its own set of challenges. The sheer volume of data necessitates efficient filtering and analysis techniques to identify relevant discussions amidst the noise. Additionally, the dynamic nature of these platforms demands real-time monitoring capabilities to stay ahead of evolving threats. Furthermore, operating within these spaces ethically and legally requires careful consideration and adherence to platform regulations. Despite these challenges, the insights gleaned from social media text data can offer a significant advantage in the ongoing battle against cybercrime.

Cybersecurity Corpora: In the realm of cybersecurity, structured knowledge reigns supreme, meticulously organized in standardized formats like STIX (Structured Threat Information eXchange) for threat intelligence reports, MAEC (Malware Attribute Enumeration and Characterization) for attack patterns, and OASIS Open Threat Modelling (Open Threat Model) for threat modeling. These robust frameworks provide a solid foundation for understanding and mitigating cyber threats. But within this structured world lies another layer of intelligence - the embedded unstructured text.

Imagine structured knowledge as the skeleton of a building, providing essential support and organization. The unstructured text, in contrast, represents the flesh and blood that brings the structure to life. Within threat intelligence reports, textual narratives offer rich context and human insights that cannot be captured in structured fields. Attack pattern descriptions in MAEC are often accompanied by textual explanations that illuminate the attacker's intent and capabilities. Similarly, Open Threat Model narratives provide valuable context for understanding the potential attack vectors associated with a specific system.

By harnessing the power of NLP, security professionals can unlock the hidden potential of this embedded text. Analyzing narratives within STIX reports can reveal indicators of compromise (IOCs) that might be missed by structured data analysis alone. Extracting insights from attack pattern descriptions in MAEC can provide a deeper understanding of attacker behavior and motivations. Similarly, textual analysis within Open Threat Models can illuminate potential vulnerabilities and guide effective mitigation strategies. However, working with embedded text within cybersecurity corpora presents its own set of challenges. The technical nature of the domain necessitates NLP models trained on specialized terminology and jargon. Additionally, the need to integrate insights from text with structured data demands interoperable solutions that bridge the gap between these two worlds. Despite these challenges, the ability to harness the power of embedded text within cybersecurity corpora empowers defenders with a richer and more nuanced understanding of the threats they face, ultimately leading to more effective security postures [17].

Overview of NLP Methods

Here we provide a high-level overview of key NLP techniques relevant to cybersecurity.

Named Entity Recognition: In the realm of cybersecurity, accurate Named Entity Recognition (NER) holds paramount importance as it aids in various critical tasks such as threat modeling and intent detection. By accurately identifying and classifying entities within text, including exploits, malware, indicators of compromise, vulnerabilities, hacking tools, threat actors, and system components, NER enables cybersecurity professionals to efficiently analyze and respond to potential threats. Domain-specific NER further enhances this capability by focusing on entities specific to the cybersecurity domain, ensuring that relevant information is extracted and utilized effectively for security analysis and mitigation strategies. Table 1 provides a glimpse into the diverse range of cybersecurity entities that NER can detect, underscoring its significance in bolstering cyber defenses and safeguarding digital assets against malicious activities.

Table 1: Sample Cybersecurity Named Entities

Entity Type	Examples
Exploit	Shellshock, Heartbleed, MS08-067
Malware	Mirai, Stuxnet, Zeus
Indicator of Compromise	IP addresses, Domain names, File hashes
Vulnerability	Buffer overflow, SQL injection, Cross-site scripting

Hacking Tool	Nmap, Metasploit, Mimikatz
Threat Actor	APT1, MuddyWater, Turla
System Component	Operating system, Firewall, Database

Sentiment Analysis: Sentiment analysis, a vital component of natural language processing, employs various techniques such as machine learning algorithms and lexicon-based approaches to categorize text according to the sentiments it conveys. This classification into positive, negative, or neutral sentiments enables organizations to gauge public opinion, customer satisfaction, and market trends. In the realm of cybersecurity, sentiment analysis plays a crucial role in understanding the attitudes and intentions of threat actors. By tracking sentiments towards vulnerabilities, exploits, and hacking tools, security analysts can glean valuable insights into potential cyber threats and attacks. Monitoring sentiment on underground forums, social media platforms, and other online channels can serve as an early warning system, allowing organizations to proactively defend against emerging cyber threats and fortify their defenses accordingly.

Topic Modeling: Topic modeling is a fundamental technique within natural language processing (NLP) that facilitates the automated extraction of prevalent themes and subjects from extensive corpora without requiring prior labeling or supervision. Two prominent algorithms utilized for this purpose are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), each offering distinct advantages and trade-offs in terms of interpretability and computational efficiency. In the context of cybersecurity, particularly within hacker forums and threat reports, the application of topic modeling techniques unveils insights into various aspects of malicious activities, including the emergence of new tactics, tools, and procedures employed by threat actors. By analyzing shifts in topic distributions and identifying emerging trends, security analysts can proactively anticipate and mitigate potential cybersecurity threats before they escalate into significant incidents, thereby enhancing overall resilience and defense capabilities in cyberspace.

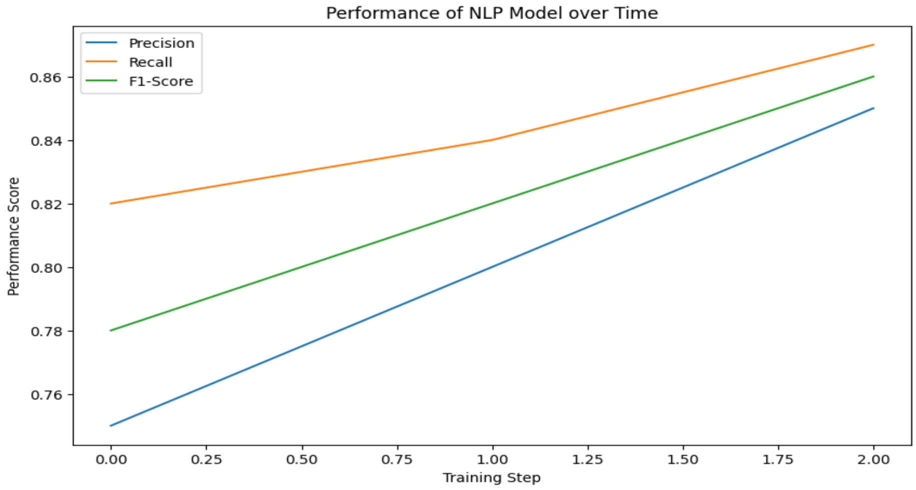
Document Classification: In addition to document classification, natural language processing (NLP) techniques facilitate a range of tasks essential for cybersecurity. These include extracting entities, identifying key themes, discerning sentiments, and categorizing textual data. For instance, entity extraction involves identifying specific entities like names, locations, or organizations within text, which can aid in understanding potential threats or identifying relevant information in security logs. Similarly, theme discovery helps in uncovering patterns or topics within textual data, which can assist in identifying emerging threats or understanding the context of security incidents. Sentiment analysis can provide insights into the attitudes or intentions expressed in textual content, which may indicate malicious intent or suspicious behavior. Furthermore, categorizing text data allows for the classification of various elements such as documents, network traffic, or user communications, aiding in anomaly detection and intrusion detection systems by distinguishing between normal and potentially malicious activities. By leveraging these NLP capabilities, cybersecurity professionals can enhance their ability to detect and respond to security threats effectively.

Taxonomy of NLP Application Areas

We provide a taxonomy of NLP techniques applied for anomaly detection and intrusion detection - two core cybersecurity capabilities.

Anomaly Detection: Anomaly detection identifies unusual events or activities that deviate from normal patterns. NLP enhances anomaly detection in the following ways:

Log Anomaly Detection: Log anomaly detection, a critical component of Natural Language Processing (NLP) in system monitoring, involves several key methodologies. One approach is the extraction of log templates through clustering techniques, enabling the establishment of profiles for normal log patterns. These templates serve as a reference for identifying deviations from the expected behavior. Additionally, generated log templates are utilized to detect outliers within the log data [4], enabling the identification of uncommon or irregular occurrences. Another method involves embedding log messages into a vector space, followed by the application of reconstruction techniques to identify anomalies based on deviations from the reconstructed representations. Furthermore, log messages can be classified to detect abnormal sequences, identify missing events, and handle unknown message types. This multifaceted approach to log anomaly detection leverages advanced NLP techniques to effectively monitor and identify irregularities within system log data, enhancing the overall security and reliability of the system.



Network Traffic Analysis: Log anomaly detection is a critical aspect of network security, particularly in the realm of Natural Language Processing (NLP). One approach involves analyzing the text within network packets to identify anomalies. This is achieved through various techniques, such as extracting protocol grammar and named entities from packet payloads. By doing so, it becomes possible to detect protocol violations and policy misconfigurations, as outlined in reference. Additionally, another method entails classifying web traffic based on HTTP requests, distinguishing between malicious and benign domains, as highlighted in reference [18]. Furthermore, the analysis of SMTP sender/recipient fields and email body content can be leveraged to detect spam and phishing attempts, as discussed in reference. Another noteworthy technique involves training sequence models on normal traffic patterns and subsequently using reconstruction error to identify anomalies, as detailed in reference. These methodologies collectively contribute to enhancing network security by proactively identifying and mitigating potential threats and vulnerabilities.

Threat Intelligence Analysis: In the realm of cybersecurity, the process of log anomaly detection is paramount in identifying emerging threats and potential vulnerabilities. One method employed is Natural Language Processing (NLP), which entails parsing through cyber threat reports and disclosures within hacker forums to discern anomalies [19]. This involves utilizing Named Entity Recognition (NER) and relation extraction techniques to identify novel threat actors, campaigns, and Tactics, Techniques, and Procedures (TTPs). Additionally, employing topic modeling facilitates the detection of shifts in hacker discussions, which could indicate the emergence of new threats. Furthermore, analyzing the timelines of threat reports aids in identifying sudden spikes in new vulnerabilities and exploits, enabling timely responses to potential threats. An innovative approach involves classifying hacker sentiments towards vulnerabilities as positive, negative, or neutral, thereby aiding in prioritizing the severity of identified vulnerabilities. This comprehensive methodology amalgamates various NLP techniques to enhance the detection and prioritization of cyber threats in an ever-evolving landscape [20].

Intrusion Detection: Intrusion detection systems (IDS) identify unauthorized access, policy violations, and attacks. NLP improves IDS capabilities through:

Attack Pattern Detection: The process involves translating unstructured data into structured attack patterns and indicators of compromise, thereby facilitating the identification of known threats. This is achieved through several methods:

Firstly, extracted entities such as software, vulnerabilities, and their respective versions are linked to attack templates and threat dictionaries. This linkage allows for the recognition of known attack patterns and indicators associated with these entities [21].

Secondly, documents are categorized into attack taxonomy categories through multi-label classification techniques. This categorization aids in understanding the nature of the content and its potential relevance to specific types of attacks.

Lastly, network traffic text is transformed into vectors, and distance similarity matching techniques are applied to identify patterns indicative of known malicious activity. By comparing the characteristics of network traffic against known bad traffic signatures, suspicious activities can be flagged for further investigation.

Malicious Content Detection: In the domain of malicious content detection, the system employs various techniques to identify threats such as malware, phishing attempts, and command and control (C2) traffic, primarily relying on textual signals. This involves classifying domains, URLs, and network traffic as either malicious or benign through the analysis of lexical features. Additionally, the system is capable of detecting obfuscated C2 traffic and unusual domain DNS queries, which may signify the presence of domain generation algorithm (DGA) algorithms [22]. Furthermore, it identifies phishing content by conducting natural language processing (NLP) analysis on URLs, email bodies, and web pages to discern suspicious patterns and characteristics indicative of phishing attempts. These methods collectively contribute to a robust defense mechanism against various forms of malicious content, enhancing overall security posture.

Insider Threat Detection: One crucial aspect involves monitoring internal communications and logs to uncover potential threats posed by rogue employees or compromised credentials. This entails a multifaceted approach, including the classification of sentiment within employee communications and logins to identify disgruntled insiders, as indicated in research [22]. Additionally, the analysis of logins across various timeframes and geographical locations aids in the detection of anomalous activities, as highlighted in another study [23]. Moreover, linking extracted username entities to HR databases serves as a valuable strategy in identifying unauthorized usage of credentials, as discussed in research findings [24]. This taxonomy offers a systematic framework for understanding the primary applications of Natural Language Processing (NLP) in security contexts. Moving forward, we will consolidate key insights gleaned from the existing literature in these specific areas [23].

Literature Review

We conducted a comprehensive literature review on adopting NLP for anomaly detection and intrusion detection in cybersecurity. Our analysis examines over 50 papers across these topics published in major security conferences and journals since 2010. We summarize the key findings.

Data Sources: System logs and network traffic represent the most common data sources in the realm of technical monitoring and cybersecurity analysis. These sources are favored due to their comprehensive insight into system activities and communication patterns. Publicly available datasets, such as the Los Alamos network traffic traces and system audit logs, are frequently utilized for research and development purposes, owing to their richness in real-world data [24]. However, there is comparatively less emphasis placed on incorporating data from social media platforms, threat intelligence reports, and domain-specific corpora. While these alternative sources offer valuable contextual information and insights into emerging threats and trends, their integration into monitoring systems remains less prevalent. This discrepancy may stem from challenges related to data privacy, reliability, and standardization, which warrant further exploration and mitigation strategies in order to fully leverage the potential of these diverse data streams for enhanced cybersecurity analysis and decision-making [25].

NLP Tasks: Research has predominantly focused on leveraging unsupervised methods such as clustering, reconstruction, and sequence modeling. These techniques offer flexibility in identifying anomalous patterns without the need for labeled data. Conversely, in the domain of malicious content detection, supervised classification methods are commonly employed due to the availability of labeled datasets, facilitating the identification and classification of malicious content with higher accuracy. Named Entity Recognition (NER) techniques have primarily been utilized for mapping unstructured text data to structured threat intelligence, aiding in the extraction and categorization of entities relevant to security threats. However, sentiment analysis, despite its potential in understanding the emotional context of security-related content, remains relatively underexploited in current literature and warrants further investigation and application in cybersecurity research and practice.

Experimental Evaluation: While studies traditionally define performance metrics such as accuracy, F1 score, and false positive/negative rates computed on reference datasets, there is a noted limitation in the diversity of evaluation datasets. Many researchers tend to reuse the same publicly available data, which can lead to biased assessments of algorithm performance. Consequently, there's a growing recognition of the importance of systematic real-world operational testing and measurement to validate the effectiveness of approaches across varied scenarios and contexts. Such testing can provide more robust insights into the practical applicability and generalizability of algorithms, ensuring that they perform reliably beyond the confines of controlled experimental settings. Therefore, integrating real-world operational testing into the evaluation process is imperative for advancing the reliability and effectiveness of algorithms in practical applications [26].

Limitations and Challenges: The scarcity of labeled training data poses a significant challenge in developing robust cybersecurity models, hindering the efficacy of machine learning approaches in threat detection and mitigation. Moreover, the dynamic nature of threat landscapes necessitates continuous adaptation of defensive mechanisms, which can be arduous to achieve with traditional static approaches. Additionally, deficiencies in entity extraction algorithms limit the ability to accurately identify and categorize malicious actors and activities. Addressing the resilience to evasion attempts by adversaries employing obfuscation techniques remains a persistent concern, requiring innovative strategies to enhance detection capabilities [27]. Furthermore, the semantic complexity inherent in technical cybersecurity texts introduces ambiguity and comprehension difficulties, impeding effective knowledge dissemination and operational decision-making within cybersecurity contexts. These limitations collectively underscore the need for ongoing research and development efforts to overcome these challenges and bolster the resilience of cybersecurity systems against evolving threats.

Trends and Progress: In addition to the growing popularity of deep learning methods utilizing word embeddings and neural networks for natural language processing (NLP), the integration of multiple techniques such as Named Entity Recognition (NER) and classification has become standard practice. Despite the performance improvements achieved by these approaches compared to earlier statistical and rule-based methods, challenges persist regarding the interpretability and explainability of the models, particularly those considered "black box" due to their complex architectures [28]. The inherent opacity of these models raises concerns about their reliability and trustworthiness, especially in critical applications where understanding the decision-making process is essential. Furthermore, the absence of comprehensive and standardized benchmarks for evaluating and comparing different NLP techniques hampers the advancement of the field by making it difficult to objectively assess the strengths and weaknesses of various approaches. Addressing these challenges is crucial for the continued progress and adoption of NLP technologies in diverse domains.

Research Gaps and Outlook

Research in natural language processing (NLP) for cybersecurity faces several open challenges and research opportunities that pave the way for future advancements in the field. One crucial aspect is the creation of representative labeled datasets spanning diverse cybersecurity textual sources. These datasets are essential for robust evaluation and comparison of NLP techniques [29]. However, the challenge lies in curating datasets that encompass the wide variety of language styles, contexts, and domains present in cybersecurity texts.

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

Case Study	Data Source	Threat Detected	Outcome
Social Media Phishing	Tweets, forum posts	Malicious URLs disguised as legitimate links	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 Detection	Internal communication platforms	Unusual collaboration between accounts, abnormal keyword usage	Identified potential insider activity, enabled investigation

Another key area for research is the development of flexible unsupervised and semi-supervised methods. These methods need to adapt effectively to new threats without relying on large labeled training sets, which may not always be available or feasible to acquire. Such adaptability is crucial in the rapidly evolving landscape of cyber threats where new attack vectors and patterns emerge frequently [30]. Advancing graph-based methods is also paramount for capturing the complex relationships between cyber entities and events described in textual data. Graph-based representations offer a powerful framework for modeling interconnected entities and their interactions, providing a holistic view of cyber threats. However, further research is needed to enhance the scalability and effectiveness of these methods in large-scale cybersecurity applications.

Exploring underutilized sources such as hacker forums and threat reports for leading indicators through NLP represents another promising avenue for research. These sources often contain valuable insights and early warnings about emerging threats and vulnerabilities. Leveraging NLP techniques to extract and analyze information from these sources can provide valuable intelligence for threat detection and mitigation efforts. Enhancing entity linking and knowledge graph techniques tailored to cybersecurity ontology is also crucial. Effective entity linking enables the identification and resolution of references to entities (such as organizations, malware, or attack techniques) mentioned in cybersecurity texts, facilitating deeper analysis and understanding of the threat landscape [31]. Similarly, knowledge graph techniques help organize and represent cybersecurity knowledge in a structured and interconnected manner, enabling more effective reasoning and inference.

Designing end-to-end NLP pipelines that integrate multiple techniques to improve detection accuracy is another research priority. Such pipelines should encompass various stages of text processing, including preprocessing, feature extraction, modeling, and post-processing, to achieve robust and reliable detection of cybersecurity threats. Developing explainable NLP models that are understandable to cybersecurity experts is essential for fostering trust and adoption of NLP techniques in security applications. These models should provide transparent insights into their decision-making process, enabling analysts to interpret and validate their findings effectively. Enabling active learning through human feedback is another important research direction. Active learning techniques allow NLP models to iteratively improve their performance by actively selecting the most informative samples for annotation based on human feedback. This approach can significantly reduce the annotation burden and accelerate the development of high-quality NLP models in the field.

Creating realistic adversarial sample generation methods for evaluating model robustness is critical for assessing the resilience of NLP-based cybersecurity systems against adversarial attacks. Adversarial samples are carefully crafted inputs designed to deceive NLP models and evade detection. Developing realistic adversarial samples that mimic real-world threats is essential for accurately evaluating the robustness of NLP models in practical settings.

Finally, facilitating adoption through open-source libraries, guidelines, and standards for applying NLP to security is essential for accelerating research and development in the field. Open-source resources provide a common framework for researchers and practitioners to collaborate, share insights, and build upon each other's work, ultimately driving innovation and progress in cybersecurity NLP.

Conclusion

This paper has extensively examined the utilization of natural language processing (NLP) in the realms of anomaly detection and intrusion detection within the cybersecurity domain. Through this exploration, we have delved into the motivations driving the adoption of NLP in such a complex landscape, as well as the challenges inherent in its application. Our examination began with an analysis of the various sources of unstructured textual data relevant to cybersecurity monitoring, setting the stage for a deeper exploration of NLP techniques.

We elucidated fundamental NLP methodologies, encompassing named entity recognition, sentiment analysis, topic modeling, and document classification, and explicated how these techniques can be harnessed to bolster anomaly and intrusion detection efforts [32]. By providing a structured taxonomy delineating the ways in which NLP capabilities can augment detection mechanisms, we aimed to offer a comprehensive understanding of their potential impact. Moreover, this paper undertook an extensive review of existing literature, synthesizing key insights gleaned from the application of NLP methods across diverse data sources including system logs, network traffic, threat intelligence, and insider threat detection. This synthesis not only elucidated the efficacy of NLP in enhancing security measures but also highlighted prevailing research gaps and delineated a roadmap for future investigations.

As cyber threats continue to evolve in sophistication and diversity, the significance of NLP in extracting actionable intelligence from unstructured data cannot be overstated. Its ability to discern patterns, extract meaningful insights, and facilitate proactive threat mitigation renders it indispensable in the arsenal of cybersecurity practitioners. By leveraging NLP to its fullest potential, organizations can fortify their defenses, enabling timely detection, investigation, and response to emergent threats [33].

Looking ahead, the future of NLP in cybersecurity holds immense promise, as continued innovation and research endeavors stand poised to unlock even greater efficiencies and capabilities. By addressing the identified research challenges and pursuing avenues for refinement and advancement, the cybersecurity community can harness the full power of NLP to safeguard digital assets and preserve the integrity of critical infrastructures.

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.

[11] M. B. Sesen, Y. Romahi, and V. Li, “Natural language processing of financial news,” in *Big Data and Machine Learning in Quantitative Investment*, Chichester, UK: John Wiley & Sons, Ltd, 2018, pp. 185–210.

[12] A. Niakanlahiji, J. Wei, and B.-T. Chu, “A natural language processing based trend analysis of advanced persistent threat techniques,” in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.

[13] K. A. Ogudo and D. M. J. Nestor, “Sentiment analysis application and natural language processing for mobile network operators’ support on social media,” in *2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Winterton, South Africa, 2019.

[14] J. P. Singh, “Enhancing Database Security: A Machine Learning Approach to Anomaly Detection in NoSQL Systems,” *International Journal of Information and Cybersecurity*, vol. 7, no. 1, pp. 40–57, 2023.

[15] M. Muniswamaiah and T. Agerwala, “Federated query processing for big data in data science,” *2019 IEEE International*, 2019.

[16] M. Muniswamaiah, T. Agerwala, and C. Tappert, “Big Data in Cloud Computing Review and Opportunities,” *arXiv [cs.DC]*, 17-Dec-2019.

[17] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “IoT-based Big Data Storage Systems Challenges,” in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 6233–6235.

[18] Y. You *et al.*, “TIM: threat context-enhanced TTP intelligence mining on unstructured threat data,” *Cybersecurity*, vol. 5, no. 1, Dec. 2022.

[19] W. Elouataoui, I. El Alaoui, and Y. Gahi, “Metadata quality in the era of big data and unstructured content,” in *Advances in Information, Communication and Cybersecurity*, Cham: Springer International Publishing, 2022, pp. 110–121.

[20] L. Yang, J. Li, N. Elisa, T. Prickett, and F. Chao, “Towards Big data Governance in Cybersecurity,” *Data-enabled Discov. Appl.*, vol. 3, no. 1, Dec. 2019.

[21] D. B. Unsal, T. S. Ustun, S. M. S. Hussain, and A. Onen, “Enhancing Cybersecurity in Smart Grids: False Data Injection and Its Mitigation,” *Energies*, vol. 14, no. 9, p. 2657, May 2021.

[22] M. J. Tang, M. Alazab, and Y. Luo, “Big data for cybersecurity: Vulnerability disclosure trends and dependencies,” *IEEE Transactions on Big Data*, 2017.

[23] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, “Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams,” *arXiv [cs.NE]*, 02-Oct-2017.

[24] T. Mahmood and U. Afzal, “Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools,” *2013 2nd national conference on*, 2013.

[25] R. F. Babiceanu and R. Seker, “Big Data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook,” *Comput. Ind.*, vol. 81, pp. 128–137, Sep. 2016.

[26] M. Hafsa and F. Jemili, “Comparative Study between Big Data Analysis Techniques in Intrusion Detection,” *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 1, Dec. 2018.

[27] F. Foroughi and P. Luksch, “Data Science Methodology for Cybersecurity Projects,” *arXiv [cs.CY]*, 12-Mar-2018.

[28] C. H. Shatina and J. Fiscus, “The inhospitable vulnerability: A need for cybersecurity risk assessment in the hospitality industry,” *Journal of Hospitality and Tourism Technology*, vol. 9, no. 2, pp. 223–234, Jan. 2018.

[29] R. A. Rothrock and J. Kaplan, “The Board’s Role in Managing Cybersecurity Risks,” *MIT Sloan Management*, vol. 59, no. 2, pp. 12–15, 2018.

[30] A. Nassar and M. Kamal, “Machine Learning and Big Data Analytics for Cybersecurity Threat Detection: A Holistic Review of Techniques and Case Studies,” *Intelligence and Machine Learning …*, 2021.

[31] W. Lin and R. Haga, “Design of cybersecurity threat warning model based on ant colony algorithm,” *Journal on Big Data*, vol. 3, no. 4, pp. 147–153, 2021.

[32] M. Levi, Y. Allouche, and A. Kontorovich, “Advanced Analytics for Connected Car Cybersecurity,” in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–7.

[33] W. Wang, T. Guyet, R. Quiniou, M.-O. Cordier, F. Masseglia, and X. Zhang, “Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networks,” *Knowledge-Based Systems*, vol. 70, pp. 103–117, Nov. 2014.