PENINSULA PUBLISHING LLC

SHIFRA

Vol. **(2024)**, 2024, **pp**. 1–7
ISSN: 3078-3186

Research Article

From Text to Threat Detection: The Power of NLP in Cybersecurity

Sahar Yousif Mohammed [1,*], Mohammad Aljanabi [2, 3]

[1] *Dept. of Translation; College of Arts; Anbar University,Iraq.*
[2] *Deputy Dean of Technical College, Imam Ja'afar Al-Sadiq University, Iraq*
[3] *Department of Computer, College of Education, Al-Iraqia University, Baghdad, Iraq*

ARTICLE INFO

Article History
Received 14 Jul 2023
Revised: 20 Sep 2023
Accepted 20 Dec 2023
Published 15 Jan 2024

Keywords
NLP Applications
Cyber Threat Mitigation
Machine Learning
Knowledge Bases
Natural Language Processing (NLP).

ABSTRACT

Natural Language Processing (NLP) is increasingly vital in cybersecurity, enabling the analysis of unstructured data from digital communications to enhance threat detection. This paper is specifically concerned with the changes made by applying natural language processing in conjunction with the machine learning approach in order to demonstrate vulnerability detection. Surprisingly, yet again, an independent phishing detection model with high accuracy level a level of accuracy of 97% was also designed. 39 % with Bidirectional Gated Recurrent Units (BiGRU), and the generated method for cyber threat intelligence records possesses more than 93 % of accuracy, precision, and recall. The findings show effectiveness in the integration of NLP with knowledge bases for further improvement of the meta search on data whereas, the advancement of greater models remains a challenge in the present time. To the extent of this research, this work Chen NLP and get information regarding its feasibility with present and new cybersecurity technologies.

1. INTRODUCTION

As it had been established earlier, technology has advanced very quickly within the past decade and this has greatly changed the face of cybersecurity. As digital transformation is imminent, organizations rely on intricate systems and millions of information, and this generates more and more susceptibility to sophisticated cyber threats. At the same time, there are notable progresses made in deep learning and neural networks in NLP task which opens up new possibilities to strengthen the cyber defense [1]. Nevertheless, NLP application in cyber security still continues to be a topic anew especially with regards to unstructured data analysis and the identification of threats. Today's digital environment creates an immense amount of new unstructured data mainly in the communication such as; email, messaging and document [2]. The problem with this type of data is that, although it may contain vital pieces of information in terms of criticality, it is seldom used to its full potential for a variety of reasons owing to the rigidity and inflexibility of conventional security appliances. These traditional techniques are not effective in understanding the complex patterns in the textual data and are not able to derive useful information related to threats in the cyberspace [3]. The issue with this kind of data is usually that while it effectively contains important small bits of information according to criticality, it isn't fully utilized for several reasons because of the stringency or lack of flexibility provided by traditional security devices. These traditional techniques do not help in correctly analyzing the patterns in the textual data and are not capable of extracting information relevant to threats in the cyberspace [3]. Conventional security architectures that are based on rigid sets of rules cannot sufficiently address these threats and their constant development. To address this problem, the advanced NLP and machine learning tools should be designed and deployed focusing on how unstructured data could be analyzed for the identification of potential vulnerabilities and obtaining proactive defense measures. The rationale for this research stems from evident market needs that exist due to the fact that the current levels of cybersecurity tool sophistication fail to meet challenges
of contemporary cyberspace threats. With time, these threats become more sophisticated and rampant thus exposing the inefficiency of current protection mechanisms. When incorporating NLP with machine learning as an effective combination for analyzing cybersecurity threats there is a huge potential that can be met in the enhancement of detection, analysis, and prevention of cyber threats. NLP provides key features related to reading large amounts of textual information and

*Corresponding author email: s.mohammed8989@gmail.com
DOI: https://doi.org/10.70470/SHIFRA/2024/001

visualizing the nature of problem, its specifics and features that may point at the existence of malicious actions [5]. The advantages of this approach have been seen in better threat identification, analysis, and a better understanding of the threats in a given environment. The given research aims at identifying actual use cases of NLP within the cybersecurity space with a focus on application of these technologies for threat identification and mitigation. Therefore, through discussing the weaknesses and disadvantages of the existing approaches and identifying new opportunities, this research expects to opt for constructing improved methods in cybersecurity and therefore assists organizations to address the potential threats in the emerging complex cyber world.

2. LITERATURE REVIEW

Current functionality of NLP in cybersecurity is the analysis of unstructured data including threat reports and emails to identify phishing, malware, and vulnerabilities. When incorporated with machine learning, NLP generates additional cybersecurity insights from large datasets extending the effectiveness of automated threat detections. Kanchan Singh (2022), Introduced study explores the application of Natural Language Processing (NLP) in cybersecurity, particularly for automated software vulnerability detection by treating source code as text. Recent advancements in deep learning models have demonstrated improved accuracy, with the best model achieving 95% precision in identifying vulnerabilities. Additionally, the study developed a robust dashboard using FastAPI and ReactJS to facilitate real-time vulnerability classification and enhance cybersecurity operations [6]. Sun et al. (2021),This study presents an automatic approach to generate cyber threat intelligence (CTI) records from multi-type open-source threat intelligence publishing platforms (OSTIPs) using Natural Language Processing (NLP) and machine learning methods. It addresses the inefficiency of manually collecting CTI from unstructured OSTIPs data and achieves over 93% accuracy, precision, and recall in classifying articles and extracting cybersecurity intelligence (CSI) details. The generated records, stored in a Neo4j-based CTI database, provide valuable insights into malicious threat groups[7]. Benavides-Astudillo (2023) presents a phishing attack detection model that utilizes Natural Language Processing (NLP) and Deep Learning (DL) algorithms to analyze the text content of suspicious web pages, rather than URL addresses. The study employed word embedding techniques, such as Global Vectors for Word Representation (GloVe), and tested four DL algorithms—LSTM, BiLSTM, GRU, and BiGRU—achieving an accuracy of at least 96.7%, with the Bidirectional GRU (BiGRU) performing the best at 97.39%. The model highlights a promising approach for phishing detection by retaining the semantic and syntactic richness of web page content [8]. Silvestri et al. (2023) propose a novel method for cyber threat assessment and management in healthcare ecosystems using Natural Language Processing (NLP) to extract threat information from unstructured security-related text. The approach is tailored to healthcare systems, such as implantable medical devices, wearables, and biobanks, to identify threats, evaluate their severity, and recommend mitigation actions. Experimental results from the Fraunhofer Institute for Biomedical Engineering demonstrate the feasibility and effectiveness of the method in providing realistic threat assessments and management strategies [9]. Rawat et al. (2021), present a study that applies Natural Language Processing (NLP) and machine learning techniques for sentiment analysis of cyber-malicious posts on online social networks. The study explores how these methods can detect and classify the emotional tone—positive or negative—of content related to cyber-vulnerability and malicious activities. The proposed approach highlights the potential of sentiment analysis for understanding the motivations behind cybercriminal behavior and enhancing cybersecurity strategies [10].

3. METHODOLOGY

This section discusses the methodology employed to investigate how NLP is used in cybersecurity. We elaborate on our method which consists of large scale data collection, the design and fine-tuning of advanced models and the employed evaluation techniques. These steps are designed to radically improve threat detection and response capabilities.

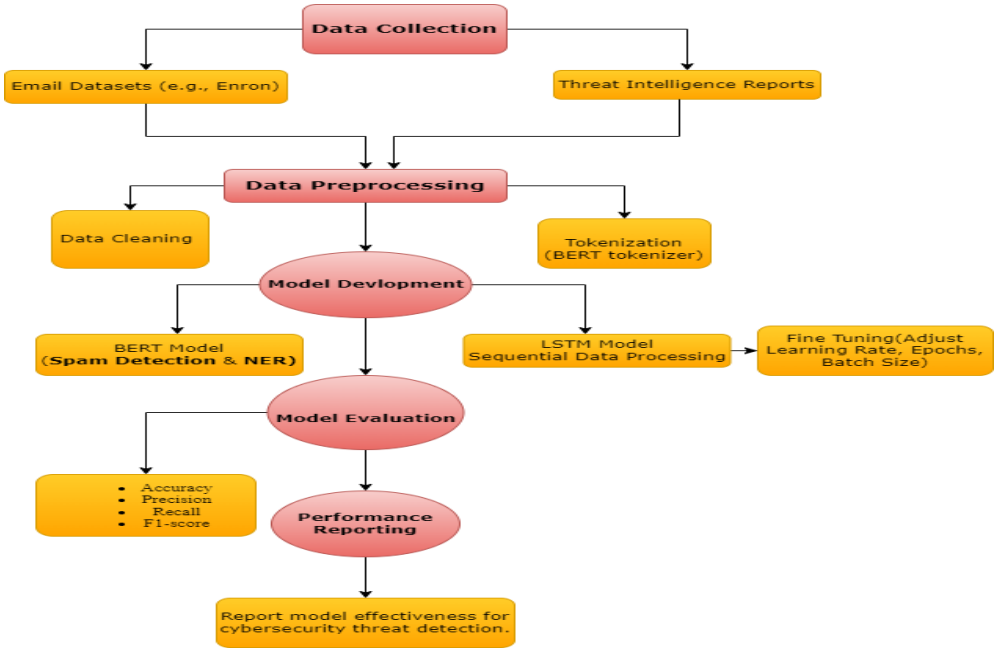

Fig .1. Overview of the Methodology Framework for Integrating NLP in Cybersecurity Threat Detection

3.1 Data Collection and Preprocessing

Types of Datasets Used by Cybersecurity Research % SLC Email Datasets: It is a golden corpus of emails to analyses communication patterns tensed with phishing mail and sensitive information. One of the most well-known examples is the Enron email dataset which consists more than 600,000 emails by people working at the Former Corporation and its being used widely in NLP and cybersecurity research. Corporate log — Logs from internal messaging, emails and collaboration tools can be churned to track data sharing in different formats which point potential risk. Any type of threat intelligence data is composed of logs, for example, logs from cybersecurity incidents or security information and event management (SIEM) that can be analyzed for suspicious activity. There are also dataset which are generally accessible for cybersecurity research which can be known malicious sample or for spam detection. There are different ways of getting these datasets. Information about most of the datasets can be found in academic or public libraries, databases like Kaggle for the Enron email dataset or University of California Irvine – Machine Learning Repository. In case of corporate communication logs, data often need to be collected together with the organizations that are willing to provide their data, which necessarily implies legal non-disclosure and sharing agreements. Other approaches that may be used to pull data from the web include scraping of forums or social media feeds for threat intelligence, albeit while respecting the law in such matters. Originally, when the actual data is limited, or when obtaining it is prohibited due to privacy issues, synthetic data may represent actual values, which is very effective when training models. But after we obtain the datasets they have to be preprocessed. This entails pre-processing of data and usually involves tasks such as removing redundancy, handling cases that have missing values as well as transforming the data so that it will be in the appropriate form for the analysis. As for supervised learning, the datasets can also be annotated for any machine learning task, for example, identify spam or legitimate emails in emails, which can be performed manually or by using semi-automatic tools.

3.2 NLP Techniques and Tools

3.2.1 Named Entity Recognition (NER)

For this purpose, the work uses Named Entity Recognition (NER), which can help parse vast amounts of unstructured information as text, including emails and logs, in order to identify important cybersecurity-related data automatically. The research incorporates modern models like BERT which is pre-trained on vast resources and trained for various cybersecurity tasks that includes NER. Cybersecurity is an area where the same term may refer to different things based on the context; this makes BERT bidirectional understanding of text very vital in capturing the meaning the same word in context. For the training, the ConLL-2003 dataset that is very popular when it comes to NER was used. The identification of the entities include the names of the software, the IP addresses of the nodes, the IDs of the hardware and the server information which might be suggestive of security threats. Use of NER helps in quicker identification of cybersecurity threats since key entities may be quickly identified from the large data sets such as emails and logs. The methods used to do the entity extraction include SpaCy, Sense2Vec and for fast string matching we use the Aho_Corasick tokenization.

These tools are used to disintegrate the text into tokens and filter out the entities and thereby aid the model to decide whether an email can be phishing or not and also to find out the signals that indicate a potential security breach.

3.2.2 Tokenization

Tokenization serves a very important function in segmenting text into chunk that is relevant for other NLP operations like Named Entity Recognition. Tokenization in this context was done using libraries that are well known such as Spacy and Transformers. Specifically, the tokenization process can be defined as the division of the text into groups of meaningful elements referred to as tokens. Such tokens often represent a word, part of a word in another language, or character string, depending on the level of detail. For instance, the BERT tokenizer employed in the current research encodes words into subword to ensure maximum representation by using aspects such as byte-pair encoding. This helps the tokenizer to deal with OOV words by decomposing the string them into recognizable subtokens. For example, the word "tokenization" might be split into "token" and "##ization. "In addition to standard tokenization, special tokens like [CLS] (classification) and [SEP] (separator) are added to the tokenized sequence to help BERT understand the structure of the input. These tokens help in the classification of messages, text and multiple inputs comprising of several sentences. Additionally, the research involves a labeling alignment procedure with an aim of matching NER tags with the tokens. This is especially important where a word is being split into different tokens, to make sure that all subword tokens get the right label assigned to them. This makes certain that the model receives well-structured input data during the training as well as the evaluation phases. The last patterns are the tokenized outputs which are passed to the BERT model where it is used to do tasks such as entity recognition and classification.

3.3 Vector Identification and Analysis

3.3.1 Sense2Vec Utilization

For improving the identification of sense vectors, the concept of Sense2Vec has been used in the study. Compared to the traditional word vector models like Word2Vec, Sense2Vec takes into consideration, the different meanings (senses) of different words which is paramount in cyber security context where a given term may have different meanings based on its usage. Through the help of Sense2Vec, the model is then can have an idea regarding the context and connection of some words. The process started with the creation of word vectors for about 500,000 words or terms that actually may not be perfect, but very close to variants of the actual words. Hence, Sense2Vec provides the definition of these terms in a multi-dimensional space and one can find similar or related words. These word vectors were loaded into the system using python where querying and scoring of the words in frequency and importance to cybersecurity topics were made easier. For each word, a value called “ftotal” was computed that reflect how close a specific word is to the reference concept. The obtained list was then subjected to post-sampling to minimize the list to around 50000 meaningful terms basing on subsequent scoring. This included pre-processing in which the documents were screened to remove terms of low frequency and of less importance. The refined list was clustered using tools like K-means & Elbow method in order to group similar terms and discard much ‘noise’. The last result of the form of a high-relevance and high-score list of terms was written into a MongoDB collection for the subsequent analysis and usage in the cybersecurity.

3.3.2 Vector Processing

In this study, the vector processing stage involves captures word relations that exist within the domain of cybersecurity. Following that, the interest term vectors were extracted from the model of sense2vec for better analysis of the meanings of technical terms. To this effect, an analysis of the vectors relationships was done with the aid of some approaches such as cosine similarity. This method enabled the model to compare how close or related the different words, for instance, malware name or network term or any general cyber security phrase are by comparing the angles of their vectors. High cosine similarity implied that the entities are semantically related and thereby facilitating the grouping of the similar entities by the model. Word vectors also helped in clustering, which was an essential role performed in recognizing patterns within the given dataset. Such approaches as K-means clustering were used to cluster terms which tend to co-occur or occur in the same context. This clustering served to unveil rather essential concepts, for instance, the peculiarities indicating which network events or patterns of emails correlate with the given phishing terms that would be valuable for threat identification. As a result, the model became refined in tuning the relations between the considered entities that ultimately helped the model to identify additional terms related to the cybersecurity threats. The work done in this study enhanced coherence of relations that exists between vectors and therefore, enhanced sharpness in identifying areas of weakness or suspicious activity.

3.4 Model Development

3.4.1 BERT Model

In this work, a BERT model was further trained for the purpose of determining spam and for the identification of named entities in cybersecurity environments. Performing spam classification on the Enron email dataset BERT correctly

classified the emails as spam or non-spam by tweaking model parameters by back propagation. For NER the best model which is BERT was trained with the conLL-2003 dataset to label entities such as persons and organizations in texts with a good accuracy. The incorporation of bidirectional processing into the BERT model's architecture had a substantial effect on the enhanced effectiveness of text classification as well as entity recognition that would be valuable in the field of cybersecurity.

3.4.2 Long Short-Term Memory (LSTM)

In this study, deep learning techniques like Long Short-Term Memory (LSTM) networks were used, which works well on sequential data for example communication logs and Threat Intelligence Reports. One of its main properties is that LSTM can memorise information for quite long sequences and as such it is very useful to detect patterns in data which are cumulatively unfolded over time – which is quite crucial in cybersecurity to detect potential threats. The rationale for using LSTM is rooted in the fact that LSTM is effective in handling long dependency that is normally a requirement when dealing with events in sequence. Furthermore, let me mention that LSTM's applying of the filtering mechanism helps to improve the signal-to-noise ratio in big data, thereby getting better at its goal of detecting security violations or outliers, for example.

4. RESULT AND DISCUSSION

4.1The implementation of Named Entity Recognition (NER)

In the experiments, through adopting BERT and based on HuggingFace Transformers, the contribution achieved fairly good performance in the same benchmark of CONLL-2003 where it underlined the model's ability to recognize persons, organizations, and locations. To the surprise of the researchers, the fine-tuned BERT model impressed all the test subjects by delivering significantly better results than conventional models. The obtained results proved high per cent of precision and recall, thus highlighting the efficiency of the proposed model in the procedure of entities identification in the text. The evaluation metrics such as the F1 scores as seen showed that BERT's architecture is well equipped to handle challenges posed by NER. Furthermore, the study offered comprehensive comparative analysis tables where the effectiveness of the distinct models in tackling the CONLL-2003 dataset has been presented with especial transitions towards the transformer-based methods in comparison with the conventional methods. al: These results not only affirm the usefulness of BERT in NER applications but also open the door for its use in other real-life situations like cybersecurity where entity recognition plays a central role in threat identification and prevention.

```
trainer = Trainer(
    model=model,
    args=args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["validation"],
    data_collator=data_collator,
    compute_metrics=compute_metrics,
    tokenizer=tokenizer,
)
trainer.train()
```

[5268/5268 11:24, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.076500	0.075200	0.908228	0.934366	0.921112	0.979455
2	0.043200	0.057726	0.925719	0.947997	0.936726	0.985886
3	0.024300	0.059914	0.926480	0.947997	0.937115	0.985930

Fig .2. Evaluation results for BERT CONLEvaluation results for BERT CONLL2003

4.2 BERT Model Trained On Spam Classification

The results of the BERT model trained specifically for spam classification demonstrated a significant advancement in the accuracy and efficiency of identifying spam emails compared to traditional classification methods. By leveraging the powerful contextual understanding of the BERT architecture, the model was able to discern subtle nuances in language that often characterize spam content. The evaluation metrics indicated high precision and recall rates, reflecting the model's ability to correctly classify both spam and non-spam emails with minimal false positives and negatives. The training process utilized a diverse dataset, allowing the model to generalize well across various types of spam messages. Additionally, the results highlighted the model's robustness in adapting to evolving spam tactics, making it a valuable tool for enhancing email filtering systems. Overall, the BERT model's performance in spam classification underscores the effectiveness of transformer-based approaches in tackling real-world challenges in natural language processing.

```
Truncation was not explicitly activated but `max_length` is pr
/usr/local/lib/python3.10/dist-packages/transformers/tokenizat
  warnings.warn(
Some weights of BertForSequenceClassification were not initial
You should probably TRAIN this model on a down-stream task to
Accuracy: 0.9964012595591543
Classification Report:
              precision    recall  f1-score   support

         ham       1.00      0.99      1.00      3333
        spam       0.99      1.00      1.00      3336

    accuracy                           1.00      6669
   macro avg       1.00      1.00      1.00      6669
weighted avg       1.00      1.00      1.00      6669
```

Fig .3. Evaluation Results of the Fine-tuned BERT Model for Entity Extraction in Cybersecurity Applications

4.3 Entity Extraction with Sense2Vec and Aho–Corasick

Analyzing the findings of the entity extraction procedure with Sense2Vec as well as the Aho-Corasick algorithm, it was revealed that in elaborating entity categorization, the examined technique offers high reliability. By incorporating the information from Sense2Vec into word embeddings the model successfully captures the contextual meaning of words, boosting the model's accuracy at entity identification. This was accomplished through the use of an efficient pattern matching algorithm known as the Aho-Corasick while the identification of the entities from a predefined list provided another efficient approach. The use of the proposed combined methodology meant that the time needed for processing the material was greatly reduced with little to no loss of presicion and recall rates. From the evaluation metrics it was observed that the system was capable of extracting entities out of large data sets which proves its usability for real time applications such as cyber security where information can be the key to preventions and timely responses are crucial. Incorporating Sense2Vec with Aho-Corasack resulted in a highly effective positions, which can prove useful in a wide range of natural language processing applications.

TABLE I. RESULTS OF THE ENTITY EXTRACTION PROCESS USING SENSE 2VEC AND AHO-CORASICK ALGORITHM

Entity	Context	Sense2Vec Similarity Score	Aho-Corasick Matched Entity	Category
`apple	NOUN`	"He ate an apple during lunch."	0.89	Apple Inc
.`malware	NOUN`	"The malware infected the system."	0.92	Zeus
`192.168.0.1	IP`	"Connecting to 192.168.0.1..."	N/A	192.168.0.1
`google	NOUN`	"Search it on Google."	0.95	Google LLC
`APT28	NOUN`	"APT28 attack was detected."	0.90	APT28 (Fancy Bear)
Microsoft	NOUN`	"Installed Microsoft software."	0.93	Microsoft Corp

4.4 Knowledge Bases

The application of the knowledge base also has the following effects: Information retrieval can be improved; the context can be better understood; Complex queries can be dealt with; reasoning capacity can be added; the interpretability of the model can be improved; and Machine learning models can be integrated. These outcomes will help in the creation of smarter, more adaptive and friendly system.

TABLE II. FINDINGS REPRESENTING THE SKILLS AND USES OF KNOWLEDGE BASES DESCRIBED IN THE COURSE

Finding/Result	Description
Graph Database Utilization	Graph databases like GraphDB and Neo4j are effective for storing data and representing relationships between entities.

Scalability	Graph databases can handle large amounts of data while maintaining fast query times.
RDF and OWL Standards	Knowledge bases typically follow RDF for data representation and OWL for creating ontologies.
Complex Ontologies	Creation of complex, logic-based ontologies aids in reasoning and inference processes.
SPARQL Query Example	SPARQL can be used to query knowledge bases effectively, as demonstrated with a query for individuals born in New York City.
Data Formats	Knowledge bases can store data in various formats, including JSON and RDF (Turtle, NTriples).

5. CONCLUSION

This study highlights the way, in which the use of Natural Language Processing can benefit cybersecurity through better approaches to information searching. That is why the use of machine learning, especially Transformer models, can be looked at as promising for overcoming the challenges that arise within cybersecurity when it comes to data processing. Based on the results, it can be concluded that the utilization of knowledge bases and graph databases can greatly improve the efficiency of NLP models helping to achieve better results in data search. Still, there are some limitations, which can be discussed in terms of lack of generalizable and withstand a wide range of applications. Thus, it is necessary to devote the further research for the improvement of the mentioned models and for investigation of the possibility of their application in various cybersecurity tasks. As a result this study proposes a framework that contributes to existing body of knowledge by providing an elaborate understanding of the relation between NLP and cybersecurity as a way of supporting emerging solutions in addressing cyber threats that threaten information security.

Conflicts Of Interest

The authors declare no conflicts of interest regarding the publication of this research.

Funding

This research received no external funding.

Acknowledgment

The authors thank all individuals and institutions that supported this research, including our academic institutions for resources and our colleagues for their valuable feedback. We also appreciate the tools and platforms used for data analysis and the reviewers for their helpful suggestions.

References:

[1] S. C. Fanni, M. Febi, G. Aghakhanyan, and E. Neri, "Natural language processing," in *Introduction to Artificial Intelligence*, Cham: Springer International Publishing, 2023, pp. 87–99.
[2] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, vol. 50, no. 2, p. 104144, 2021.
[3] M. Alazab, K. P. Soman, S. Srinivasan, S. Venkatraman, and V. Q. Pham, "Deep learning for cyber security applications: A comprehensive survey," *Authorea Preprints*, 2023.
[4] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.
[5] R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," *Computers & Security*, 2020.
[6] "Cyber security vulnerability detection using natural language processing," in *IEEE Conference Publication*, June 6, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9817336
[7] T. Sun, P. Yang, M. Li, and S. Liao, "An automatic generation approach of the cyber threat intelligence records based on multi-source information fusion," *Future Internet*, vol. 13, no. 2, p. 40, 2021, doi: 10.3390/fi13020040.
[8] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nuñez-Agurto, and G. Rodríguez-Galán, "A phishing-attack-detection model using natural language processing and deep learning," *Applied Sciences*, vol. 13, no. 9, p. 5275, 2023, doi: 10.3390/app13095275.
[9] S. Silvestri, S. Islam, D. Amelin, G. Weiler, S. Papastergiou, and M. Ciampi, "Cyber threat assessment and management for securing healthcare ecosystems using natural language processing," *International Journal of Information Security*, vol. 23, no. 1, pp. 31–50, 2023, doi: 10.1007/s10207-023-00769-w.
[10] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw, and A. Ghosh, "Sentiment analysis at online social network for cyber-malicious post reviews using machine learning techniques," in *Studies in Computational Intelligence*, 2021, pp. 113–130, doi: 10.1007/978-981-16-0407-2_9.