International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 395-399, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE64
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Automated Cyber Threat Identification Using Natural Language Processing

[1]Parumanchala Bhaskar, [2]Farooq Sunar Mahammad, [3]K. Ramachari, [4]S.Arbas Basha, [5]K.Vivekananda Reddy, [6]B.Malleswara Reddy, [7]S.Ravi Teja

[1,2,3,4,5,6,7]Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, 518501, India
E-mail: [1]bhaskar.cse@srecnandyal.edu.in, [2]farooq.cse@srecnandyal.edu.in

***Abstract* - This abstract challenge by leveraging Natural Language Processing (NLP) to automate cyber threat identification. The proposed system utilizes advanced NLP techniques to analyse vast amounts of textual data from sources such as cybersecurity reports, social media, forums, and dark web communications. The proliferation of cyberthreats in today's digital world poses serious security and privacy issues. Because malevolent behaviour is dynamic, traditional threat detection techniques usually fall behind. By increasing threat detection's precision, effectiveness, and scalability, the solution seeks to strengthen digital infrastructures' resistance to cyberattacks. Because the digital world moves quickly, cyberattacks are becoming more sophisticated and larger than ever before. By developing a mechanised system for cyber threat identification using Natural Language Processing (NLP), this research will address this basic problem. By improving threat detection's precision, effectiveness, and scalability, the solution aims to strengthen digital infrastructures' defences against cyberattacks.**

Keywords: Intelligence on Cyber Threats, Cybersecurity, Natural Language Processing (NLP), Automated Threat Recognition, Analysis of Threats, Machine Learning, Deep Learning, Information Mining, Extraction of Information.

I. INTRODUCTION

The Problem: We have tons of information about potential cyber threats. This information is scattered everywhere – in security logs, reports, social media, etc. It's too much for humans to handle, and it's hard to find the real threats quickly. Plus, the threats are always changing.

The Solution: NLP can help! Think of NLP as teaching computers to understand human language. Instead of just seeing words, computers can actually understand the meaning behind the words.

How NLP Helps: Understands the Data: NLP can read all that messy threat information (logs, reports, posts) and make sense of it. It's like having a super-fast reader that understands what it's reading.

Finds Patterns: NLP can spot hidden patterns and clues that might suggest a cyber-attack is happening or about to happen. It can find connections that humans might miss.

NLP helps computers understand and analyse huge amounts of text data to find cyber threats faster and more effectively than humans can alone. It's like giving cybersecurity experts a powerful assistant that can read and understand everything!

II. PROBLEM STATEMENT

Cybercriminals are using new vulnerabilities in computer systems very quickly after they are discovered. This gives defenders very little time to protect against attacks. The main contribution of our work is the approach to characterize or profile the identified threats in terms of their intentions or goals, providing additional context on the threat and avenues for mitigation.

2.1 Project Goals

This project aims to develop an automated framework for cyber threat identification using NLP, focusing on real-time analysis and efficient threat detection. The specific objectives include:

- Creating an automated pipeline for data collection, preprocessing, and analysis.
- Constructing NLP models capable of classifying and detecting threats in real-time.
- Evaluating the performance and scalability of the automated framework.

2.2 Scope

This document focuses on the development of an automated NLP-driven system for identifying cyber threats. The range includes:

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 395-399, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE64
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

- Automated analysis of text data from various sources (e. g., emails, social media, security logs).
- Deployment of real-time threat detection functionalities.
- Assessment of the system's performance regarding accuracy and processing speed.

III. LITERATURE REVIEW

The Internet is Important but Risky: We rely on the internet for everything, but it's also full of dangers like cyberattacks. Threat Intelligence Helps: To fight these attacks, we use "threat intelligence". This is like gathering clues about upcoming attacks, including details regarding the attacker's methods ("signatures"). This prepares us. Where We Obtain Clues: We gather these clues from different places:

Formal Sources: Authorized organizations that share threat information in a methodical, organized way (like a formal report).

Informal Sources: More casual sources, such as news articles, blogs,, discussions.

Organized Clues are Ideal: When the clues are structured ("organized"), security tools can more readily understand them and take automated measures to keep us protected.

In summary: We collect signs of cyberattacks from multiple sources. The more organized these pieces of information are, the more effectively we can shield ourselves. However, the slide indicates that there remains a significant amount of unstructured, chaotic information that is challenging to utilize, and that's where the new danger arises.

IV. EXISTING RESEARCH

Current research has examined multiple aspects of NLP based cyber threat detection, including:

- Automated phishing identification using machine learning techniques.
- Real-time threat surveillance on social media utilizing NLP strategies.

Automated extraction of threat intelligence from security documents. Implementation of message queuing and stream processing for handling large data volumes. Studies have also explored the use of NLP models in cloud environments to achieve scalability and efficiency.

4.1 Drawback in Existing System:

Contextual Ambiguity: Natural language is heavily reliant on context and can be ambiguous. The same word or phrase may carry different meanings based on the context, making it challenging to predict and characterize emerging cyber threats.

Limited Generalization: NLP models often do not generalize well across different fields, sectors, or languages. A model that is trained on a specific category of data may not perform optimally in another scenario.

Explainability and Trust: NLP models are frequently perceived as black boxes, making their decision-making processes difficult to understand. This lack of clarity can undermine trust and hinder widespread use.

Adversarial Attacks: NLP models are susceptible to adversarial attacks, where malicious individuals deliberately manipulate input data to deceive the model.

V. PROPOSED SYSTEM

The framework coordinates three fundamental elements: First, the identification of cyber threats and their classification; second, the profiling of these identified threats, distinguishing their motives and goals through a sophisticated machine learning architecture; and third, the issuance of alerts based on the danger posed by the identified threats. A significant innovation in our work lies in our approach to define these emerging threats, providing contextual understanding of their motives. This improved layer of understanding not only enhances threat detection but also offers avenues for effective countermeasures. In our experimental research, the profiling stage achieved an impressive F1 score of 77%, demonstrating a strong ability to identify and understand identified threats." "This Paper leads the forefront of proactive cybersecurity strategies, aiming to equip defenders with a sophisticated system capable of performing early threat detection and advanced threat characterization. By utilizing a rich source of event data and advanced machine learning techniques, the framework not only identifies threats but also delves deeper into their motives, providing valuable insights for proactive defence strategies against rapidly evolving cyber threats.

5.1 Advantages of Proposed System:

Early Threat Detection: NLP can analyse vast amounts of text-based data from various sources in real-time, thereby enabling early identification of new and developing cyber threats. Early detection allows organizations to take proactive measures before the threats escalate.

Flexibility to Evolving Threats: NLP models can be consistently trained and updated to keep up with changing cyber threats. By incorporating new data and retraining the models periodically, the system can stay current and effectively identify emerging threats.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 395-399, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE64
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Improved Situational Awareness: By processing and interpreting natural language, NLP systems enhance situational awareness. Organizations can comprehend the shifting threat landscape, learn about adversarial tactics, and make informed decisions concerning cybersecurity measures.

Cost-Efficiency: NLP-based automation of threat detection reduces the reliance on manual labour for analysing large volumes of text data. This cost-effective approach allows organizations to allocate resources more efficiently and focus on strategic cybersecurity initiatives.

VI. METHODOLOGY

Data Collection and Preprocessing:

An automated pipeline will be created to gather text data from various sources. Data preprocessing will be automated, including: Text cleaning, tokenization, and normalization. Feature extraction will utilize techniques like TF-IDF and word embeddings. Data streams will be processed using message queueing technologies.

NLP Model Development and Deployment: "NLP models, including both machine learning models and deep learning models, will be created to classify and identify threats. The models will be optimized for real-time processing in terms of latency and throughput. The models will be launched in a scalable environment, such as a cloud-based platform or a containerized setting. "

Real-Time Threat Detection: A real-time threat detection system will be implemented, utilizing stream processing technologies. The system will continuously analyse incoming data streams and trigger alerts for detected threats.

Evaluation Metrics: Accuracy, precision, recall, F1-score, latency, throughput, and scalability metrics.

VII. EXPECTED OUTCOMES

This Paper aims to:

- Develop an automated framework for realtime cyber threat identification.
- Enhance the efficiency and effectiveness of threat detection through automation.
- Provide a scalable and adaptable solution for large-scale data analysis.
- Improve the speed of response to cyber threats.

Future research will focus on:

- Integrating automated threat response capabilities.
- Improving the explainability of real-time threat detection models.
- Implementing adaptive learning techniques for continuous model improvement.
- Building a report to future screening of identified threats.

VIII. RESULT

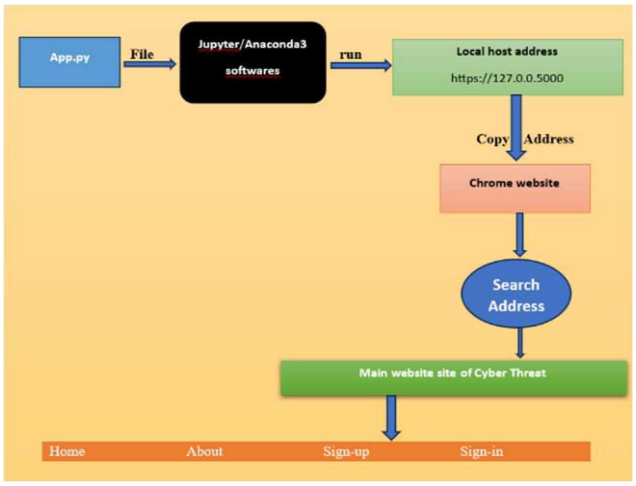

Figure 1: Architecture

Figure 2: Cyber Threat Not Detected

Figure 3: Cyber Threat Detected

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 395-399, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE64
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

IX. CONCLUSION

We researched a range of machine learning algorithms - Naive Bayes, SVM, KNN, Random Forest, Bagging, Boosting, Neural Networks, and Voting Classifier—in this research that are specifically well-suited for various classification and prediction tasks. These algorithms find important applications across a range of fields from text classification and image recognition to anomaly detection and ensemble learning. What makes them work is their capability to manage complex data relationships, respond to differing datasets, and make accurate predictions.

In the future, machine learning algorithms for applications such as cybersecurity, health diagnostics, and autonomous systems have promising prospects. Developments in deep learning and reinforcement learning are poised to make algorithms even better, allowing more advanced applications in real-world situations. Additionally, combining these algorithms with new technologies like edge computing and quantum computing could provide new paths for processing speed and accuracy.

REFERENCES

[1] Mahammad, Farooq Sunar, et al. "Key distribution scheme for preventing key reinstallation attack in wireless networks." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

[2] Suman, Jami Venkata, et al. "Leveraging natural language processing in conversational AI agents to improve healthcare security." Conversational Artificial Intelligence (2024): 699-711.

[3] Sunar, Mahammad Farooq, and V. Madhu Viswanatham. "A fast approach to encrypt and decrypt of video streams for secure channel transmission." World Review of Science, Technology and Sustainable Development 14.1 (2018): 11-28.

[4] Mahammad, Farooq Sunar, Karthik Balasubramanian, and T. Sudhakar Babu. "Comprehensive research on video imaging techniques." All Open Access, Bronze (2019).

[5] Mahammad, Farooq Sunar, and V. Madhu Viswanatham. "Performance analysis of data compression algorithms for heterogeneous architecture through parallel approach." The Journal of Supercomputing 76.4 (2020): 2275-2288.

[6] Devi, M. Sharmila, et al. "Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language." Journal of Research Publication and Reviews 4.4 (2023): 497-502.

[7] Devi, M. Sharmila, et al. "Machine Learning Based Classification and Clustering Analysis of Efficiency of Exercise Against Covid-19 Infection." Journal of Algebraic Statistics 13.3 (2022): 112-117.

[8] Mandalapu, Sharmila Devi, et al. "Rainfall prediction using machine learning." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

[9] Chaitanya, V. Lakshmi, et al. "Identification of traffic sign boards and voice assistance system for driving." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

[10] Chaitanya, V. Lakshmi. "Machine Learning Based Predictive Model for Data Fusion Based Intruder Alert System." journal of algebraic statistics 13.2 (2022): 2477-2483.

[11] Chaitanya, V. Lakshmi, and G. Vijaya Bhaskar. "Apriori vs Genetic algorithms for Identifying Frequent Item Sets." International journal of Innovative Research & Development 3.6 (2014): 249-254.

[12] Parumanchala Bhaskar, et al. "Incorporating Deep Learning Techniques to Estimate the Damage of Cars During the Accidents" AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

[13] Parumanchala Bhaskar, et al "Cloud Computing Network in Remote Sensing-Based Climate Detection Using Machine Learning Algorithms" remote sensing in earth systems sciences (springer).

[14] Parumanchala Bhaskar, et al. "Machine Learning Based Predictive Model for Closed Loop Air Filtering System." Journal of Algebraic Statistics 13.3 (2022): 416-423.

[15] Paradesi Subba Rao, "Detecting malicious Twitter bots using machine learning" AIP Conf. Proc. 3028, 020073 (2024), https://doi.org/10.1063/5.0212693.

[16] Paradesi Subba Rao,"Morphed Image Detection using Structural Similarity Index Measure"M6 Volume 48 Issue 4 (December 2024), https://powertechjournal.com.

[17] Mr.M.Amareswara Kumar,EFFECTIVE FEATURE ENGINEERING TECHNIQUE FOR HEART DISEASE PREDICTION WITH MACHINE LEARNING" in International Journal of Engineering & Science Research, Volume 14, Issue 2, April-2024 with ISSN 2277-2685.

[18] Mr.M.Amareswara Kumar, "Baby care warning system based on IoT and GSM to prevent leaving a child in a parked car" in International Conference on Emerging Trends in Electronics and Communication Engineering - 2023, API Proceedings July-2024.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 395-399, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE64
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Citation of this Article:

Parumanchala Bhaskar, Farooq Sunar Mahammad, K. Ramachari, S.Arbas Basha, K.Vivekananda Reddy, B.Malleswara Reddy, & S.Ravi Teja. (2025). Automated Cyber Threat Identification Using Natural Language Processing. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 395-399. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE64

****

