JDEBM
Journal of Digital Engineering and Business Managemet

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

ADVANCED CYBER THREAT IDENTIFICATION AND PROFILING USING TEXT ANALYTICS AND NLP

[#1]**Anitha Padigapati,***Assistant Professor,Dept of CSE,*
[#2]**Garlapati Swetha,***Assistant Professor,Dept of CSE,*
[#3]**Mrs.K.Nagalatha,***Assistant Professor,Dept of CSE,*
[#4]**B. Sanjana,***Student, Dept of CSE,*
[#5]**G.Aakanksha,***Student, Dept of CSE,*
[#1-5]***Scient Institute Of Technology(Autonomous), Ibrahimpatnam, R.R.Dist, TG, India.***

ABSTRACT: This research employs Natural Language Processing (NLP) to automatically identify and categorize emerging cyber risks in order to improve cybersecurity intelligence and early attack detection. Due to the quick development of digital technology and online communication platforms, cyber threats are constantly expanding and becoming more complicated. Real-time threat identification is challenging since modern threat detection systems rely on organized data and predefined signatures. Large volumes of unstructured textual data from blogs, forums, threat intelligence feeds, cybercrime reports, and social media platforms are evaluated using natural language processing (NLP) in order to address this problem. Data collection, threat entity identification, and cyber intrusion pattern detection are all done automatically by the application. The system classifies emerging cyber threats by characteristics, attack potential, and severity using machine learning and text mining. By increasing the speed and accuracy of cyber threat intelligence, this automated solution helps businesses get ready for emerging security threats. The results of the test indicate that the NLP-based method can identify and characterize distinct cyberthreats. This enhances risk reduction and safety monitoring.

Keywords:*Cyber Threat Intelligence, Natural Language Processing, Emerging Threat Detection, Text Mining, Cybersecurity, Machine Learning, Threat Profiling, Automated Security Analysis.*

1. INTRODUCTION

As digital technology, cloud computing, and interconnected networks advance, cyberspace is become increasingly complex. The frequency and complexity of cyberthreats have increased as a result of businesses using digital infrastructure for data storage, communication, and service delivery. The speed and complexity of cyber threats make traditional security methods, such as signature-based detection and human analysis, ineffective. In order to promptly detect and evaluate cyberthreats, there is a growing need for sophisticated, automated systems.

Natural language processing, or NLP, has emerged as a potent technology in recent years for extracting valuable information from vast volumes of unstructured text data. Blogs, social networking sites, hacker collectives, security bulletins, and vulnerability databases can all provide information on new cyberthreats. The complexity of this text data prevents me from independently evaluating it. NLP enables computers to identify, categorize, and comprehend hazardous data. Large

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

volumes of data can be swiftly and effectively evaluated by defensive systems.

NLP can detect possible cyberthreats on its own by examining language patterns that point to security flaws. Techniques including sentiment analysis, named entity identification, topic modeling, and text categorization are uāsed to identify critical indicators, such as malware names, attack strategies, vulnerabilities, and impacted systems. NLP-based systems can anticipate cyberattacks by evaluating threat data and online communications on their own. Businesses and security experts can obtain information more quickly with this tactic.

Explaining new hacks is crucial to comprehending their objectives and dangers. Cyber event text reports can be categorized by industry, attack kind, region, and threat actor behavior using NLP algorithms. Cybersecurity teams can arrange threat trend collections with the use of automatic profiling. As a result, students will gain the ability to identify dangers, react to situations, and develop defenses.

2.PROPOSED SYSTEM

The primary aim of this research is to design a method for automatically detecting and profiling newly emerging cyber threats using Open Source Intelligence (OSINT). The system is intended to provide cybersecurity professionals with timely alerts. The proposed framework can be summarized in the following major steps:

- Ongoing surveillance and data gathering from influential individuals and organizations on Twitter to uncover unfamiliar terms linked to cyber threats or malicious operations.
- Applying Natural Language Processing (NLP) and Machine Learning (ML) techniques to filter and classify these terms, distinguishing those that are likely to represent threat identifiers from irrelevant ones.
- Mapping discovered threats to MITRE ATT&CK tactics by analyzing procedural examples to infer the most probable techniques being employed.
- Issuing real-time alerts about newly detected or evolving threats, including their description, objectives, and an associated risk score that reflects the speed of their progression since discovery.

Advantages

When orchestrating a cyber-attack, adversaries generally need to:

- Pinpoint exploitable weaknesses,
- Obtain the tools and expertise required to leverage those weaknesses,
- Select a target and enlist collaborators,
- Build or acquire the necessary infrastructure, and
- Strategize and carry out the operation.

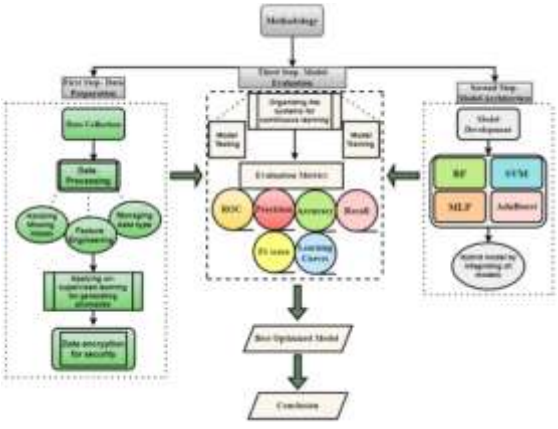

Figure1: “Proposed Cyber Threat Detection Methodology”

Decision Tree Classifiers: Due to their ease of use and simplicity, decision tree classifiers are widely used. They use a tree-like architecture to arrange data into

JDEBM
Journal of Digital Engineering and Business Management
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

branches based on criteria. A class or option is represented by each leaf node. Every path is a result, and every internal node is a test. The training data is divided repeatedly until every point in a group is classified in order to construct the model.

Gradient Boosting: Gradient boosting is a sophisticated machine learning approach for classification and regression. A resilient prediction model is produced by combining several weak models, typically decision trees. The models in this series aim to outperform their predecessors. This gradual improvement produces remarkable performance and accuracy, often outperforming competing techniques like random forests.

Logistic Regression Classifiers: The link between a categorical dependent variable and one or more independent factors is examined using logistic regression. Although it can handle a wide range of categories, it is primarily utilized in 0/1 or Yes/No scenarios. This approach is adaptable and doesn't rely on presumptions about data distribution. It is helpful for categorization tasks since it can generate success metrics, model correctness, and prediction probabilities.

Naïve Bayes: For directed learning, the Naïve Bayes algorithm makes different assumptions. This concept has a lot of real-world uses despite its simplicity. It is fast, simple to set up, and capable of managing massive amounts of data. This technique enhances content classification and spam detection. Although this may not be evident, its accuracy frequently matches that of more sophisticated equipment.

Random Forest: With many decision trees, random forest learning increases prediction accuracy. To avoid overfitting, a random combination of qualities and data are used in the construction of each tree. Regression tasks are determined by the average prediction, whereas classification problems are decided by the majority vote from all trees. Because it works with a variety of data types and yields reliable results with minimal change, it is frequently utilized.

SVM (Support Vector Machine): Support vector machines are good at classifying items. It accomplishes this by determining the optimal classification strategy for the dataset. It achieves good class separation when used with multidimensional data. Instead of modeling possible possibilities, SVM classifies by precise separation. It is distinguished by its accuracy, robustness, and ability to solve challenging classification tasks with limited data.

3. LITERATURE SURVEY

Anderson et al. (2025): Robust Natural Language Processing (NLP) technologies are used to process unstructured threat intelligence data in order to automatically identify cyber threats. Semantic trends in security reports, blogs, and dark web sources are found using transformer-based designs. It builds comprehensive profiles and automatically categorizes new threats. This facilitates the prompt identification and resolution of problems in dynamic online environments.

Fernandez & Gupta (2024): In order to detect dangerous textual activity, this study presents an NLP-driven cyber threat profile system that makes use of named entity identification and contextual embedding models. Finding compromise indicators and elucidating the connections between threat actors, tools, and vulnerabilities are the methods. It is easier to uncover zero-day vulnerabilities and

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

obtain critical data, according to experiments.

Iyer& Choudhury (2023): The study looks into whether NLP models based on deep learning can automatically identify novel hacking threats. The technology scans cybersecurity literature for hidden patterns and novel attack techniques. The method creates dynamic hazard profiles by connecting malware, attack tactics, and target systems. This helps you become more conscious of your situation and protect yourself before an attack.

Peterson et al. (2022): Neural language representations and topic modeling are used by a hybrid NLP system to detect and evaluate cyberthreats. The program connects concealed individuals to attack patterns using danger data. It helps cybersecurity experts comprehend sophisticated attacks and automatically classifies threats.

Kumar & Banerjee (2021): Using natural language processing, this cyber threat intelligence study independently finds and characterizes new risks. Sequence learning organizes danger summaries by identifying patterns in textual input. Tests show that it is simpler to identify new risks and protect against new cyberattack strategies.

4.MODULES

Service Provider:

A working Service Provider account and password are necessary for this module. Verification of user identities, training and evaluation of profile data sets, observation of training and testing results, access to all profile identity predictions, calculation and analysis of the prediction ratio, evaluation of the results, downloading of predicted data sets, and monitoring of remote users are all possible. In order to confirm training and user profile data, he can also use a bar chart.

View and Authorize Users:

The administrator has access to a complete list of every person who has registered for this feature. The supervisor is aware of the user's identity, email address, and home address. Access to the server can be facilitated by the administrator.

Remote User

There are n people in this module. Before acting, you must register. A database contains the registration data for users. He needs to provide a working login and password after registering. After logging in, users can review their profile, verify usage, and register.

5. RESULTS

Fig5.1 User login

Fig5.2 Dataset Trained and Tested Results

Fig5.3 Bar Graph

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

Fig5.4 Line Chart

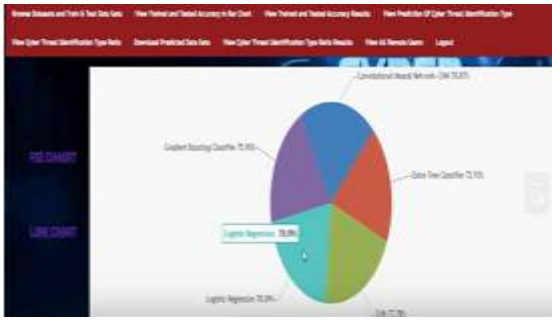

Fig5.5 Pie Chart

Fig5.6 Pie Chart

Fig5.7 Prediction of cyber threat type

6. CONCLUSION

Natural Language Processing (NLP) to autonomously discover and classify cyber threats is a novel and proactive computer security strategy. NLP methods analyze large amounts of unstructured data from security reports, social media, vulnerability databases, and hacker forums to identify dangers and early warning signals. Threat intelligence is improved by spotting patterns, extracting important entities, and creating real-time threat profiles. Organizations may better anticipate, understand, and respond to complex and pervasive cyber threats by integrating natural language processing with cybersecurity frameworks and machine learning. Digital resilience and security will improve.

REFERENCES

[1]. B. D. Le, G. Wang, M. Nasim, and A.Babar,‘‘Gathering cyberthreat intelligence from Twitter using noveltyclassification,’’ 2019, arXiv:1907.01755.

[2]. Definition: Threat Intelligence,Gartner Research, Stamford, CO, USA, 2013.

[3]. R. D. Steele, ‘‘Open source intelligence: What is it? why is it important to the military,’’ Journal, vol. 17, no. 1, pp. 35–41, 1996.

[4]. C. Sabottke, O. Suciu, and T. Dumitras, ‘‘Vulnerability disclosure in the age of social media: ExploitingTwitterforpredictingreal-worldexploits,’’ in Proc. 24th USENIX Secur. Symp. (USENIX Secur.), 2015, pp. 10411056.

[5]. A. Sapienza, A. Bessi, S. Damodaran,P. Shakarian, K. Lerman, and E. Ferrara, ‘‘Early warnings of cyber threats inonline discussions,’’ in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW),Nov. 2017, pp. 667–674.

JDEBM
Journal of Digital Engineering and Business Management
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/02
DOI: https://doi.org/10.65713/jdebmv2i202

[6]. E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, ‘‘Darknet and deepnet mining for proactive cybersecurity threatintelligence,’’ in Proc. IEEE Conf. Intell.Secur. Informat. (ISI), Sep. 2016, pp. 712.

[7]. S. Mittal, P. K. Das, V. Mulwad, A.Joshi, and T. Finin, ‘‘CyberTwitter: Using Twitter to generate alerts forcybersecurity threats and vulnerabilities,’’in Proc. IEEE/ACM Int. Conf. Adv.Social Netw. Anal. Mining (ASONAM),Aug. 2016, pp. 860–867.

[8]. K. K. Gajula and A. T. Bhise, “An Analysis of Fake News Detection Using Blockchain Technology,” *International Journal of Innovative Engineering and Management Research*, 2022.

[9]. A. Attarwala, S. Dimitrov, and A.Obeidi, ‘‘How efficient is Twitter: Predicting 2012U.S.presidentialelections using support vector machinevia Twitter and comparing against Iowaelectronic markets,’’ in Proc. Intell. Syst.Conf. (IntelliSys), Sep. 2017, pp. 646652.

[10]. K. K. Gajula, Y. K. Sharma, and R. Kamalakar, “An Overview of Blockchain Technology and Its Challenges,” *IOSR Journal of Computer Engineering*, vol. 21, no. 3, pp. 40–45, 2019.

[11]. O. Oh, M. Agrawal, and H.R.Rao, ‘‘Information control andterrorism: Tracking the Mumbai terrorist attack through Twitter,’’ Inf. Syst. Frontiers, vol. 13, no. 1, pp. 33–43, Mar. 2011.

[12]. T. Sakaki, M. Okazaki, and Y. Matsuo, ‘‘Earthquake shakes Twitter users: Real-time event detection by social sensors,’’ in Proc. 19th Int. Conf.World Wide Web, Apr. 2010, pp. 851–860.

[13]. B. De Longueville, R. S. Smith, and G.Luraschi, ‘‘‘OMG, from here, I can see the flames!’: A use case of mining location based social networks to acquire spatio-temporal data on forest fires,’’ in Proc. Int. Workshop Location Based Social Netw., Nov. 2009, pp. 73–80.

[14]. K. K. Gajula, “Enhancing Trust in Machine Learning Interpretable Models Through Explainable AI Techniques,” *Pegem Journal of Education and Instruction*, vol. 13, no. 4, pp. 909–915, 2023.

[15]. K. K. Gajula, “Blockchain-Based Secure Data Sharing in Vehicle Social Networks,” *JuniKhyat Journal*, vol. 12, no. 1, pp. 217–223, 2022.

[16]. M. K. Srinivasan and K. K. Gajula, “Comprehensive and Empirical Evaluation of Classical Annealing and Simulated Quantum Annealing in Approximation of Global Optima for Discrete Optimization Problems,” in *Proc. ICTIS*, 2021, pp. 165–181.

[17]. K. K. Gajula and K. Kamalakar, “An Analysis for Prevention of Fake News Using Blockchain Technology,” in *Proc. National Conf. on Recent Advancements on Computer Science (CONRACS)*, 2019.