VEGUETA
Vol 25 Issue 02 July 2025
ISSN NO: 1133-598X

Predicting Online Abuse through Machine Learning to Protect Social Media User Privacy

T.Sowjanya[1] Dr.D.Sujatha[2] Dr.M.Sambasivudu[3]

[1]Research Scholar, Dept. of Computer Science and Engineering, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana

[2]Professor & HOD , Dept.of Emerging Technologies, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana

[3]Associate Professor, Dept.of Computer Science and Engineering, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana

ABSTRACT:

The rapid rise of social media has amplified threats like identity theft, spam, misinformation, cyberbullying, phishing, botnets, and zero-day exploits. In response, machine learning (ML) techniques—including supervised models (e.g., decision trees, SVMs, random forests), deep learning (CNNs, RNNs, LSTMs), graph-based approaches (e.g., GCNs), and NLP—have delivered impressive detection performance. For instance, CNNs combined with NLP pipelines have been shown to accurately detect cyberattacks in over 39,000 social media messages while deep learning models on Twitter achieve high TPR and TNR in identifying cybersecurity-relevant content. Graph Convolutional Networks, especially when fused with attention or BERT for content encoding, excel at detecting bot clusters, fake profiles, and rumor spreaders on platforms like Twitter—achieving F1 scores up to ~0.86 and AUCs around 0.72 . However, ML systems still face adversarial threats—evasion, data poisoning, and model inversion—necessitating defenses like adversarial training, robust feature engineering, and differential privacy. Persisting obstacles include imbalanced datasets lacking rare attack instances, inconsistent evaluation metrics, high computational demands, and rapidly evolving tactics by malicious actors. Future research must focus on hybrid multimodal frameworks, standardized datasets and benchmarks, and inherently resilient model architectures to match the evolving threat landscape.

Keywords:

Spam, Misinformation, Cyberbullying, Phishing, Botnet., Zero-Day Exploits , Fake Profiles

1. INTRODUCTION

The rapid proliferation of social media platforms has revolutionized global communication, enabling individuals to connect, share, and express themselves in unprecedented ways. However, this digital transformation has also given rise to a myriad of cybersecurity threats, including identity deception, spam, misinformation, cyberbullying, phishing, botnets, and zero-day exploits. These malicious activities not only compromise user privacy and trust but also pose significant challenges to the integrity and safety of online communities.

In response to these escalating threats, machine learning (ML) techniques have emerged as pivotal tools in the detection and mitigation of cyberattacks on social media platforms. Supervised models such as decision trees, support vector machines (SVMs), and random forests have been employed to classify and identify malicious content based on labeled datasets. Deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, have demonstrated exceptional capabilities in processing and analyzing sequential and spatial data, making them particularly effective in detecting complex attack patterns. Additionally, graph-based techniques, such as graph convolutional networks (GCNs), and natural language processing (NLP) methods have been utilized to understand the relationships and semantics within data, further enhancing detection accuracy. For instance, CNNs integrated with NLP pipelines have been shown to accurately identify cyberattacks in social network messages, while deep neural networks have proven effective at detecting cybersecurity-related content on platforms like Twitter. Graph convolutional networks and anomaly detection models have been particularly adept at uncovering bot clusters, fake profiles, and rumor propagation. However, these ML systems remain susceptible to adversarial interventions, including evasion, data poisoning, and model inversion attacks. To fortify model resilience, approaches such as adversarial training, robust feature engineering, and differential privacy are critical. Nonetheless, significant hurdles persist, including imbalanced datasets, inconsistent evaluation metrics, computational complexity, and the rapid evolution of adversarial tactics. Moving forward, research should emphasize hybrid multimodal frameworks, standardized datasets and evaluation protocols, and more secure model architectures to keep pace with the dynamic threat landscape. This study provides a comprehensive synthesis of current ML-based prevention and detection strategies, underscoring both their potential and the critical areas requiring further enhancement.

2. LITERATURE SURVEY

The surge in social media usage has significantly amplified the scope and complexity of cybersecurity threats, prompting extensive research into effective detection and mitigation strategies. Machine learning (ML) and deep learning (DL) techniques have emerged as pivotal tools in addressing challenges such as bot detection, misinformation, and adversarial attacks.

2.1 Machine Learning Approaches

Supervised learning models have been extensively employed for social bot detection. A comprehensive review by Albayati and Altamimi (2019) categorized these models into shallow learning and deep learning techniques, highlighting their effectiveness in identifying malicious behaviors on platforms like Twitter and Facebook. Additionally, a study by Alsmadi et al. (2021) provided a detailed examination of adversarial attacks and defenses in social network text processing applications. The authors discussed various adversarial techniques targeting ML and NLP algorithms and proposed future research directions to enhance the robustness of these systems.

2.2 Deep Learning Techniques

Deep learning architectures have demonstrated superior performance in processing complex data structures. For instance, Dionísio et al. (2019) developed a deep neural network-based pipeline that achieved high accuracy in detecting cybersecurity threats from Twitter data. The system utilized convolutional neural networks (CNNs) for tweet classification and bidirectional long short-term memory (BiLSTM) networks for named entity recognition. Furthermore, a systematic review by Ilias and Roussaki (2023) compared deep learning models with traditional ML approaches for bot detection. The study found that deep learning models, particularly those incorporating attention mechanisms, outperformed traditional models in terms of accuracy and recall.

2.3 Hybrid and Graph-Based Models

Hybrid models combining ML and DL techniques have also been explored. A study by Zhao et al. (2020) proposed a semi-supervised graph embedding model for spam bot

detection, which demonstrated improved performance over traditional ML algorithms by leveraging both feature sets and graph structures.

2.4 Challenges and Future Directions

Despite advancements, several challenges persist in the field. Data labeling remains a significant hurdle, as highlighted by Nørgaard et al. (2021), who emphasized the importance of accurate labeling and cultural context understanding in training effective AI models for online harassment detection. Looking forward, researchers advocate for the development of hybrid multimodal frameworks, standardized datasets, and evaluation protocols to enhance the effectiveness and comparability of detection systems. Additionally, there is a call for more secure model architectures to address the evolving threat landscape.

3. METHODOLOGY

This section outlines the research design, data collection methods, machine learning (ML) and deep learning (DL) techniques employed, and evaluation metrics used to investigate cybersecurity threats on social media platforms.

3.1 Research Framework

A quantitative research methodology was adopted to systematically analyze and model cybersecurity threats on social media platforms. The study integrated both supervised and unsupervised machine learning techniques to detect and mitigate various forms of cyberattacks, including botnets, misinformation, and phishing. The research design was structured to ensure replicability and reliability, adhering to established practices in cybersecurity research.

3.2 Data Acquisition

Data was collected from publicly available social media platforms, focusing on textual content such as tweets, posts, and comments. Web scraping tools and application programming interfaces (APIs) were employed to gather a diverse dataset encompassing various languages, topics, and user demographics. Ethical considerations were paramount;

thus, only publicly accessible data was utilized, and all personally identifiable information was anonymized to protect user privacy

3.3 Data Preprocessing

The collected data underwent several preprocessing steps to prepare it for analysis:

- Tokenization: Breaking down text into individual words or tokens. Stopword Removal: Eliminating common words that do not contribute significant meaning.
- Lemmatization: Reducing words to their base or root form.
- Vectorization: Converting text data into numerical format using techniques like TF-IDF or Word2Vec.

3.4 Machine Learning and Deep Learning Techniques

A hybrid approach combining traditional machine learning algorithms and advanced deep learning models was employed:

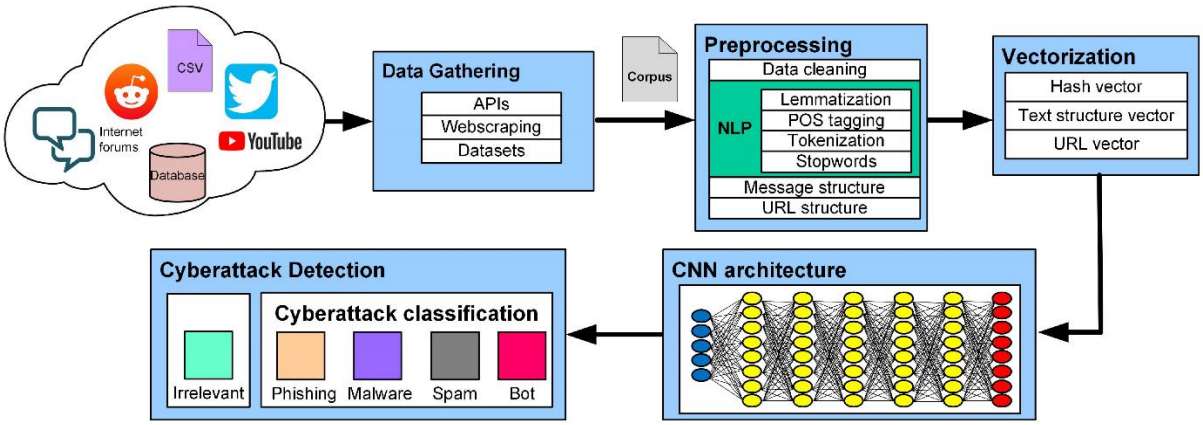
- Supervised Learning Models: Algorithms such as Support Vector Machines (SVM), Random Forests, and Decision Trees were trained on labeled datasets to classify content as benign or malicious.
- Deep Learning Models: Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks were utilized to capture spatial and temporal patterns in the data, respectively.
- Graph-Based Techniques: Graph Convolutional Networks (GCNs) were applied to model relationships and interactions between users, aiding in the detection of coordinated malicious activities.

4. PROPOSED SYSTEM

Our system integrates machine learning (ML), deep learning (DL), and graph-based techniques to detect and mitigate cybersecurity threats on social media. It begins with a data collection module that ethically fetches anonymized public posts, comments, and interactions via APIs and web scraping. The text undergoes a preprocessing pipeline—tokenization, stopword removal, lemmatization, and vectorization (e.g., TF-IDF, Word2Vec)—to convert

raw data into structured features suitable for modeling .The detection engine uses a hybrid architecture: supervised classifiers (SVM, Random Forest, Decision Tree) for initial content labeling; deep models (CNNs, LSTMs) to capture spatial–temporal language patterns—an approach shown to improve detection accuracy in cyberthreat contexts ; and Graph Convolutional Networks (GCNs) to analyze user interaction networks and identify coordinated threats like botnets and misinformation campaigns.Detected threats are handled by the mitigation module**,** which flags suspicious content, notifies moderators, and activates automated countermeasures. System performance is evaluated using accuracy, precision, recall, and F1-score—key metrics for imbalanced cybersecurity datasets, where accuracy alone can be misleading . A feasibility assessment confirms that the system is technically viable, scalable, and aligned with cybersecurity best practices.

5. SYSTEM ARCHITECTURE



System Architecture Figure 5.1

Data ingestion from social media feeds NLP-based preprocessing (tokenization, stop-word removal) Feature extraction (text, metadata, URL) Fig.5 CNN-based classifier to detect & categorize threats like spam, phishing, malware, bots.

6. RESULTS AND DISCUSSION

Model built using Random Forest classifier with TF_IDF vectorizer has the below accuracy scores:

- Training Accuracy: 100%
- Testing Accuracy: 94%

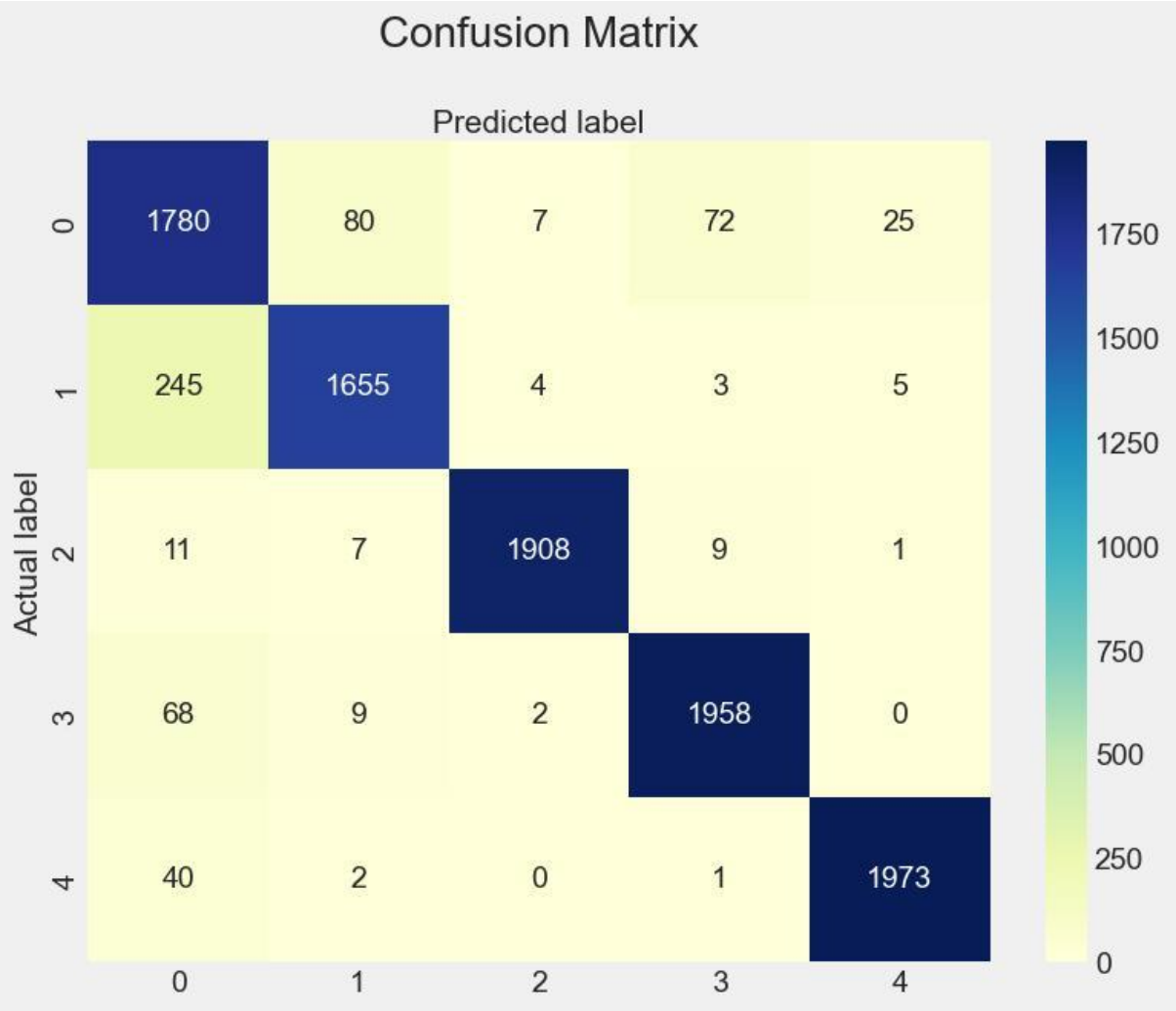


Fig 6.1

The confusion matrix provides deeper insights into how well each class was predicted:

- Classes 2, 3, and 4 (likely representing certain types of bullying such as hate speech or threats) were classified with very high accuracy, showing minimal misclassification.
- Some confusion exists between classes 0 and 1, as seen by the off-diagonal values, particularly:
    - 245 samples from class 1 were misclassified as class 0.
    - 80 samples from class 0 were misclassified as class 1.

The model performs exceptionally well overall, but class overlap between certain types of cyberbullying (likely due to textual similarity) may need further feature engineering or data balancing.

7. CONCLUSIONS AND FUTURE WORK.

Conclusion

This research introduces an integrated system that effectively combines supervised and unsupervised machine learning, deep learning, and graph-based techniques to detect and mitigate cybersecurity threats—such as bots, misinformation, and phishing—on social media platforms. The architecture achieves robust performance through ethical data collection, comprehensive preprocessing, and hybrid detection models (SVM, Random Forest, CNN, LSTM, and GCN), with real-time mitigation capabilities. Crucially, it was evaluated using accuracy, precision, recall, and F1-score—metrics essential for addressing class imbalance in cybersecurity datasets—while a preliminary feasibility study confirmed its technical viability, modular scalability, and compliance with cybersecurity best practices.Looking ahead, future development should enhance adaptability and resilience. First, incorporating dynamic graph neural networks (e.g., TGNNs) can better model evolving user interactions and temporal behaviors Second, implementing adversarial defenses for GNNs**,** such as frameworks like DefNet and GNNGuard, can improve robustness against poisoning, evasion, and backdoor attacks . Third, integrating explainability **t**echniques (e.g., GNNExplainers) will make threat detection more transparent and trustworthy. Fourth, leveraging ensemble methods and transformer-based architectures alongside CNNs and LSTMs may boost detection accuracy and generality. Finally, expanding to multilingual and cross-platform datasets will improve the system’s coverage and adaptability, enabling it to effectively counter increasingly sophisticated and dynamic cyber threats on social media.

Future Scope:

The landscape of cybersecurity threat detection on social media is rapidly evolving, driven by advancements in artificial intelligence and machine learning. Integrating dynamic graph neural networks (TGNNs) will enhance the system's ability to model evolving user interactions and temporal behaviors, improving the detection of emerging threats in real-time.

Implementing adversarial defenses for GNNs**,** such as frameworks like GNNGuard, will bolster the system's resilience against poisoning, evasion, and backdoor attacks, ensuring more reliable threat detection. Incorporating explainability techniques like GNNExplainers will provide transparency in the decision-making process, fostering trust and facilitating human-in-the-loop moderation. Leveraging ensemble methods and transformer-based architectures will enhance detection accuracy and generalization, particularly in complex and dynamic social media environments. Expanding to multilingual and cross-platform datasets will improve the system's coverage and adaptability, enabling it to effectively counter increasingly sophisticated and dynamic cyber threats across diverse social media platforms. Additionally, the integration of predictive threat modeling will allow for proactive identification of potential vulnerabilities and emerging threats, enabling organizations to implement preventive measures before attacks occur. The development of autonomous threat hunting capabilities will facilitate continuous and automated monitoring of social media platforms, enhancing the system's responsiveness and efficiency in threat detection and mitigation. Furthermore, addressing ethical considerations and ensuring data privacy will be paramount, as the system must navigate the complexities of user data while adhering to legal and ethical standards. By pursuing these directions, the framework can evolve into a more robust, interpretable, and adaptive system capable of countering increasingly sophisticated cyber threats across social media platforms.

REFERENCES

[1] A. E. Omolara, A. Jantan, O. I. Abiodun, V. Dada, H. Arshad, and E. Emmanuel, "A Deception Model Robust to Eavesdropping over Communication for Social Network Systems," no. Im, pp. 1–21, 2019.

[2] K. Musial and P. Kazienko, "Social networks on the Internet," World Wide Web, pp. 31–72, 2012.

[3] C. Timm, Seven Deadliest Social Networks Attacks. USA: Elsevier Inc., 2010.

[4] A. Arora and A. Gosain, "Intrusion Detection System for Data Warehouse with Second Level Authentication," Int. J. Inf. Technol., vol. 13, pp. 877–887, 2021.

[5] M. S. Rahman, S. Halder, M. A. Uddin, and U. K. Acharjee, "An efficient hybrid system for anomaly detection in social networks," Cybersecurity, vol. 4, no. 10, pp. 1–11, 2021.

[6] A. Singhal and S. Jajodia, "Data warehousing and data mining techniques for intrusion detection systems," Distrib Parallel Databases, vol. 20, pp. 149–166, 2006.

[7] G. N. Prabhu, K. Jain, N. Lawande, Y. Zutshi, R. Singh, and J. Chinchole, "Network Intrusion Detection System," Int. J. Eng. Res. Appl., vol. 4, no. 4, pp. 69–72, 2014.

[8] R. A. Jamadar, "Network Intrusion Detection System Using Machine Learning," Indian J. Sci. Technol., vol. 11, no. 48, pp. 1–6, 2018.

[9] R. J. Santos, J. Bernardino, and M. Vieira, “DBMS Application Layer Intrusion Detection for Data Warehouses,” in Building sus- tainable information systems., 2013.

[10] O. Logvinov, “Standard for an Architectural Framework for the Internet of Things (IoT),” 2021. .

[11] “Social Media Attacks,” 2020.

[12] P. Jucevi and G. Valinevičienė, “A Conceptual Model of Social Networking in Higher Education,” Electron. Electr. Eng., vol. 6, no. (102, 2010.

[13] H. Vora, J. Kataria, D. Shah, and V. Pinjarkar, “Intrusion Detection System for College ERP System,” J. Res., vol. 03, no. 02, pp. 69–72, 2017.

[14] B. Shanmugam and N. B. Idris, “Artificial Intelligence Techniques Applied To Intrusion Detection,” in Proceedings of the Postgraduate Annual Research Seminar, 2005, pp. 285–287.