e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:03/March-2025 Impact Factor- 8.187 www.irjmets.com

WEB APP OF MALICIOUS POST DETECTION OF SOCIAL MEDIA

Mrunali Kate[*1], Anushka Amale[*2], Shreya Ramatkar[*3], Kajal Yerone[*4], Vedanti Kadu[*5], Dr. V.S Gulhane[*6]

[*1,2,3,4,5]UG Student Department Of Information Technology Sipna C.O.E.T. Amravati, India.

[*6]Head Of Department Of Information Technology Sipna C.O.E.T. Amravati, India.

DOI: https://www.doi.org/10.56726/IRJMETS70168

ABSTRACT

In the digital age, social media platforms serve as vital communication channels but are also susceptible to the dissemination of malicious content, including hate speech, misinformation, and cyberbullying. This paper presents a novel detection system designed to identify and classify malicious posts across various social media platforms. Utilizing advanced machine learning algorithms, natural language processing (NLP) techniques, and a comprehensive dataset of labeled posts, the system effectively analyzes text features, user behavior, and contextual information to detect harmful content in real-time. The proposed system achieves high accuracy and low false-positive rates, demonstrating its potential as a reliable tool for enhancing user safety and promoting a healthier online environment.

Moreover, the system incorporates a feedback loop, allowing continuous learning from new data and evolving malicious patterns. This adaptive capability ensures that the detection methods remain effective against emerging threats. By automating the detection process, this system not only aids platform moderators but also contributes to ongoing research in combating online toxicity and fostering positive digital interactions. Ultimately, our work aims to establish a safer online ecosystem, encouraging constructive dialogue and reducing the prevalence of harmful content.

Keywords: Malicious Content, Machine Learning, Real-time Detection, Natural Language Processing (NLP), URL Detection, Text Sentiments.

I. INTRODUCTION

The rapid proliferation of social media platforms has transformed the way individuals communicate, share information, and connect with one another. While these platforms offer significant benefits, such as enhanced connectivity and the democratization of information, they also present challenges related to the spread of malicious content. Posts that promote hate speech, misinformation, cyberbullying, and other harmful behaviors can have serious implications for users' mental health and societal cohesion.

Recent studies indicate that malicious content on social media can lead to real-world consequences, including violence, discrimination, and erosion of trust in legitimate information sources. The challenge of detecting such posts is compounded by the sheer volume of user-generated content and the nuanced nature of language, which can often obscure harmful intent. Traditional moderation methods, relying heavily on human oversight, are insufficient to address this scale and complexity.

To combat these issues, there is a growing need for automated systems capable of accurately identifying and flagging malicious posts in real time. The use of machine learning and natural language processing techniques has emerged as a promising solution, enabling the analysis of vast amounts of data for patterns indicative of malicious behavior. However, developing a robust detection system poses several challenges, including the need for diverse training datasets, the ability to adapt to evolving language trends, and minimizing false positives that can undermine user trust.

This paper introduces a sophisticated detection system that leverages advanced algorithms to analyze and classify social media content. Our approach not only focuses on text-based features but also considers user interactions and contextual factors that contribute to the identification of malicious posts. By harnessing these technologies, we aim to enhance user safety, support platform moderators, and contribute to a more positive online environment.

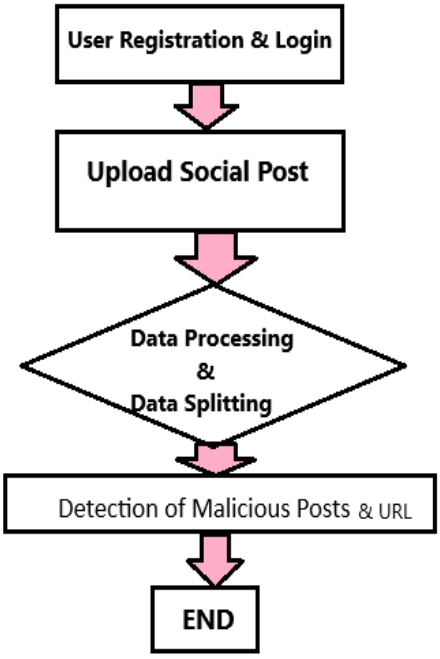
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:03/March-2025 Impact Factor- 8.187 www.irjmets.com

Furthermore, our system is designed to adapt dynamically to emerging threats by incorporating feedback mechanisms that allow it to learn from new instances of malicious content. This adaptability is crucial in a landscape where the tactics used by malicious actors are constantly evolving. By utilizing a combination of supervised and unsupervised learning techniques, we aim to improve the detection accuracy while reducing reliance on extensive manual labeling.

In addition, we address ethical considerations, ensuring that the system upholds user privacy and avoids biases that could disproportionately affect specific communities. The integration of diverse perspectives in the training phase will help create a more equitable detection model. The subsequent sections of this paper will detail the methodology, including data collection, feature extraction, and model training, followed by an evaluation of the system's performance against established benchmarks. Through our research, we hope to provide valuable insights into the ongoing fight against online toxicity and the promotion of healthy digital interactions. Ultimately, our goal is to foster a safer online community where users can engage without fear of harassment or misinformation.

II. METHODOLOGY

The diagram outlines an malicious post detection of social media system comprising three main stages: Input, Classification of dataset and ML algorithm and Detection of Social post. Typical identity deception detection experimental workflow. Data are selected from public sources or are simulated. Features are selected based on social and/or user behavior data. A model is constructed based on a machine learning algorithm, and it is further evaluated using metrics such as the f-score.



The diagram represents a process flow for a Malicious URL Detection System. It begins with User Registration & Login, where administrators create accounts to access the system. This step ensures that only authorized users can manage and oversee the detection process, maintaining the integrity of the system.

Next, in the Data Collection phase, the system continuously monitors incoming URLs from various sources, including user submissions, emails, and web pages. This phase aggregates a diverse dataset of URLs for analysis, providing a comprehensive view of potential threats.

This multifaceted approach allows for accurate detection of malicious URLs, including phishing attempts, malware distribution, and spam.

In the Review & Verification phase, administrators assess the flagged URLs to confirm their malicious status. This manual review process is essential for ensuring the reliability of the detection system and making informed decisions about necessary actions, such as adding confirmed malicious URLs to a blocklist.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:03/March-2025 Impact Factor- 8.187 www.irjmets.com

Finally, in the Feedback & Reporting stage, the system generates comprehensive reports summarizing detected malicious URLs.

In summary, the Malicious URL Detection System employs a systematic approach to identify and manage harmful URLs effectively. By integrating data collection, URL analysis, and user feedback mechanisms, the system aims to enhance online safety and provide administrators with the necessary tools to combat cyber threats proactively.

III. ADVANTAGES AND CHALLENGES

The proposed detection system offers several advantages. Firstly, it enables real-time detection of harmful content, allowing for immediate intervention to mitigate its impact. The use of advanced machine learning algorithms and natural language processing techniques results in high accuracy, minimizing false positives and ensuring that legitimate content is not wrongly flagged. Additionally, the system's adaptability is enhanced by a feedback loop mechanism, which allows it to continuously learn from new data and evolve in response to emerging threats.

Moreover, the system significantly enhances user safety by effectively identifying and classifying malicious content, thus promoting a healthier online environment. It also supports platform moderators by automating the detection process, enabling them to focus on more critical issues. The system contributes to ongoing research in combating online toxicity, providing valuable insights that can be applied across various contexts. Ultimately, by reducing harmful content, the system encourages constructive dialogue and fosters positive interactions among users.

However, there are notable challenges associated with the implementation of this detection system. One major issue is the constantly evolving nature of malicious patterns, which can complicate the system's ability to accurately identify new forms of harmful content. Additionally, achieving high contextual understanding remains a challenge, as language nuances can significantly alter meaning based on context.

Data quality and bias are also critical concerns; the system's effectiveness relies heavily on the quality of the training dataset, and biases within this data can lead to skewed results. Privacy concerns arise from monitoring user-generated content, particularly regarding data collection practices and user consent. Furthermore, deploying a robust detection system requires significant computational resources, which may pose a barrier for smaller platforms.

The risks of false positives and negatives present another challenge; while striving for high accuracy, there remains a possibility of incorrectly flagging innocent content or failing to detect harmful content. Additionally, integrating the detection system with existing platform infrastructure may entail technical challenges and necessitate substantial workflow adjustments. Finally, user resistance to perceived censorship or over-moderation can hinder the acceptance of automated detection systems.

Addressing these challenges is essential for the effective implementation and long-term success of the detection system in creating a safer online ecosystem.

IV. CONCLUSION

In conclusion, malicious post detection on social media is a critical aspect of ensuring a safe and secure online environment for users. As the volume of content shared on social platforms grows, so does the complexity of identifying harmful posts, including hate speech, cyberbullying, misinformation, and malicious links. Advanced machine learning algorithms, natural language processing, and real-time monitoring systems are pivotal in detecting these harmful activities with increasing accuracy. The future of this field holds great promise, with innovations such as multimodal content analysis, behavior-based detection, and cross-platform monitoring. However, as detection technologies evolve, they must also address challenges related to privacy, user freedom, and the ethical implications of automated moderation. Ultimately, a balanced approach that combines technological advancements with user empowerment and regulatory support will be essential to effectively combat malicious content while preserving a positive social media experience.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:03/March-2025 Impact Factor- 8.187 www.irjmets.com

V. REFERENCES

[1] H. Tarpara, K. Choudhary, T. Shah, P. Shukla and N. Kumar Chaudhary, "Post-Recovery Sanitization of Compromised Social Media Accounts," 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-6, doi: 10.1109/APCIT62007.2024.10673541.

[2] S. Liang, W. Lin, D. Li and Y. Liu, "A Method for Detecting Post-Exploitation Malicious Communication Traffic Based on Hypergraph Neural Networks," 2024 Sixth International Conference on Next Generation Data-driven Networks (NGDN), Shenyang, China, 2024, pp. 205-210, doi: 10.1109/NGDN61651.2024.10744095.

[3] S. -W. Huang, J. -L. Wu and Y. -H. Wu, "The Social Stage of Responses: Social Intent Detection in Discussion Threads Using Deep Learning Model," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8, doi: 10.1109/IJCNN60899.2024.10650297.

[4] A. K. M. Rubaiyat Reza Habib, E. Elijah Akpan, B. Ghosh and I. K. Dutta, "Techniques to Detect Fake Profiles on Social Media Using the New Age Algorithms - A Survey," 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2024, pp. 0329-0335, doi: 10.1109/CCWC60891.2024.10427620.

[5] Kagan, G. F. Alpert and M. Fire, "Zooming Into Video Conferencing Privacy," in IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 933-944, Feb. 2024, doi: 10.1109/TCSS.2022.3231987.

[6] Wang, "Analysis and detection of low quality information in social networks," 2014 IEEE 30th International Conference on Data Engineering Workshops, Chicago, IL, USA, 2014, pp. 350-354, doi: 10.1109/ICDEW.2014.6818354.

[7] N. S. Gawale and N. N. Patil, "Real Time Detection System for Malicious URLs," 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 2014, pp. 856-860, doi: 10.1109/CICN.2014.181

[8] H. S. Panchal, N. Kumar Jadav, P. Chaturvedi, R. Gupta and S. Tanwar, "Image-based Ransomware Detection for Social Media Post Using Structural Similarity Index," 2023 IEEE 20th India Council International Conference (INDICON), Hyderabad, India, 2023, pp. 649-654, doi: 10.1109/INDICON59947.2023.10440734.

[9] A. Messai, Z. F. Hamida, A. Drif and S. Giordano, "Multi-input BiLSTM deep learning model for social bot detection," 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS), BLIDA, Algeria, 2023, pp. 1-6, doi: 10.1109/ICAECCS56710.2023.10104646.

[10] Sudhir et al., "A Machine Learning Approach to Spam Detection in Social Media Feeds," 2023 IEEE 33rd International Conference on Microelectronics (MIEL), Nis, Serbia, 2023, pp. 1-6, doi: 10.1109/MIEL58498.2023.10315788.