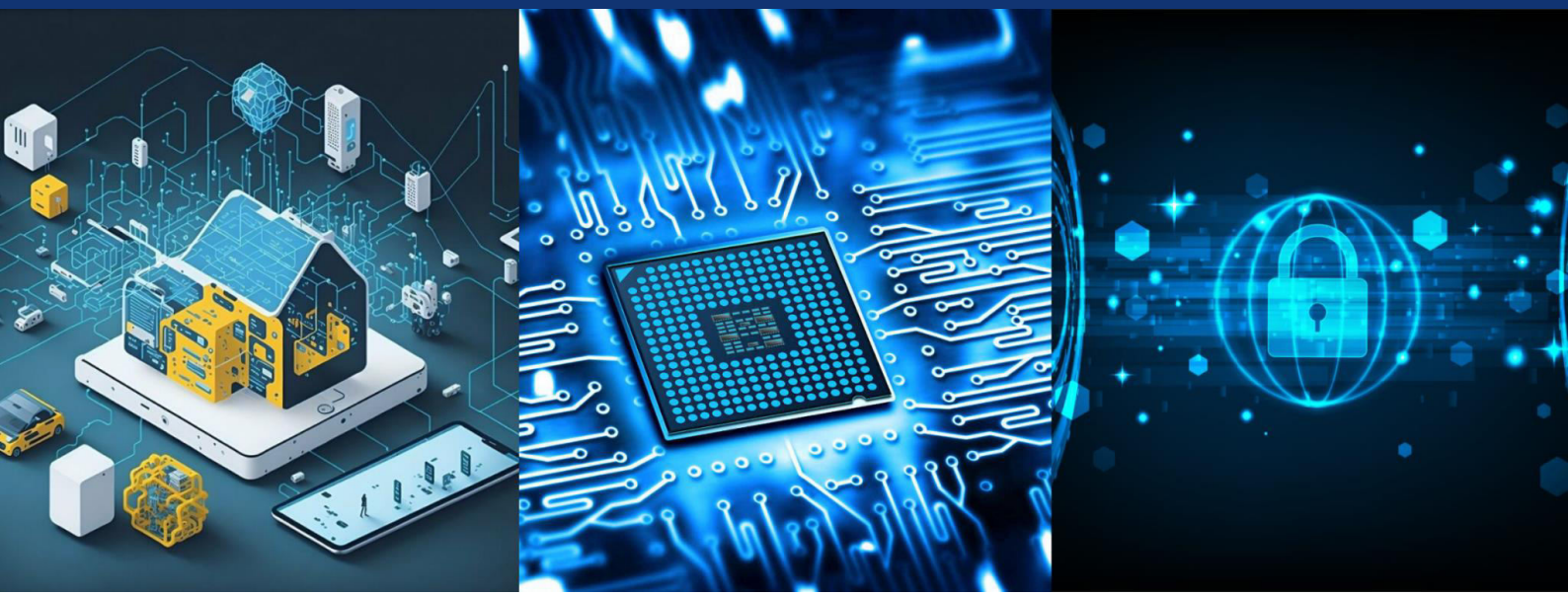
IJIRCCE

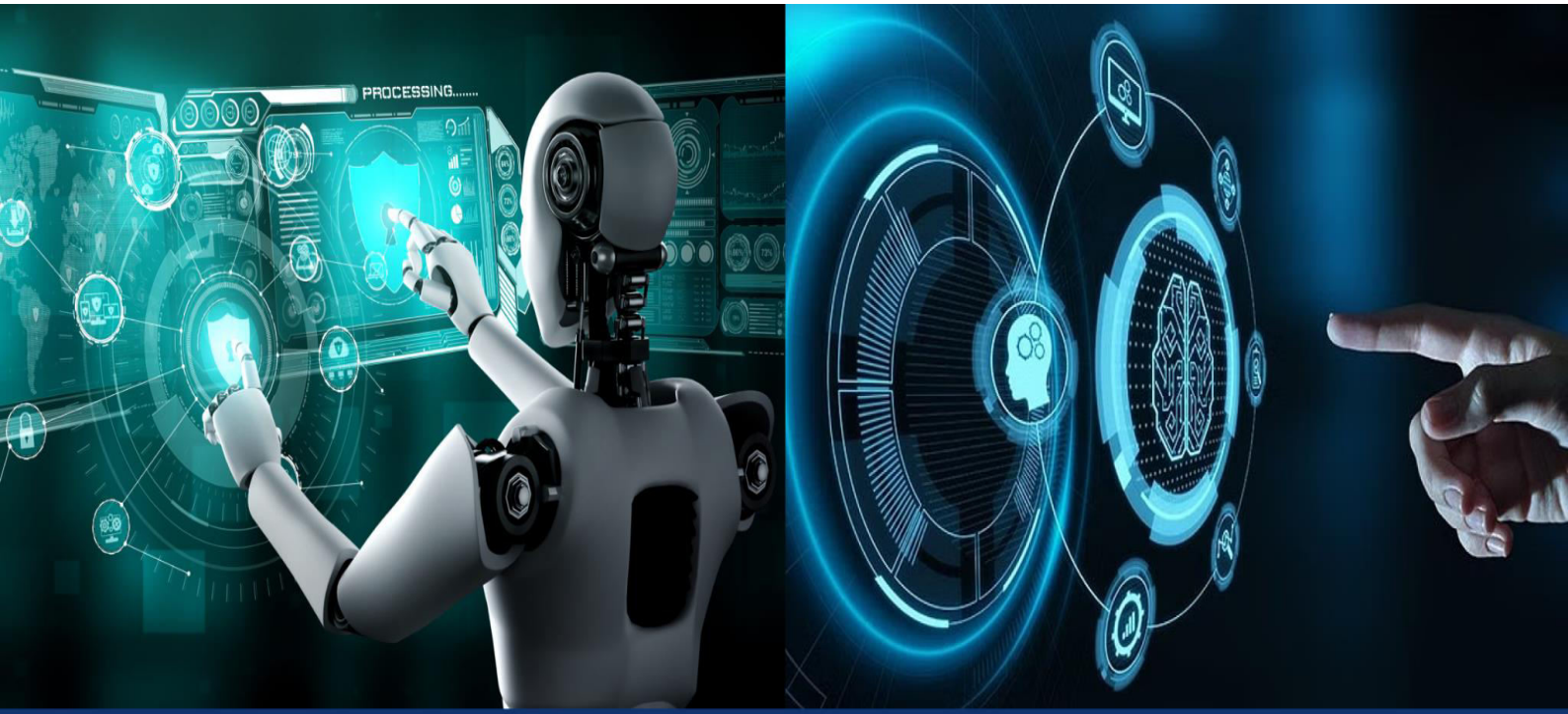
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304080

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Detection of Cyberbullying on Social Media using Machine Learning

Mr.G.Raju, Neha Kumari, S.Meghalatha, S.Madhav, T.Narsimha Reddy

Guide, Department of Computer Science and Engineering, School of Engineering, Malla Reddy University, Hyderabad, India

B. Tech, School of Engineering, Malla Reddy University, Hyderabad, India

ABSTRACT: Cyberbullying is a growing concern on social media, affecting mental health and online safety. Manual content moderation is challenging due to the vast amount of user-generated content. This project leverages machine learning (ML) and natural language processing (NLP) techniques to automatically detect cyberbullying in social media posts, comments, and messages. Using text preprocessing, feature extraction and supervised ML models , as well as deep learning approaches the system classifies text as bullying or non-bullying. Sentiment analysis and toxicity detection further enhance accuracy by identifying hate speech and offensive content. The workflow involves data collection, feature extraction, model training, classification, and automated reporting. This project has real-world applications in social media moderation, online gaming chat filtering, educational institutions, and parental monitoring. Challenges such as slang variations, sarcasm detection, and privacy concerns are addressed using advanced NLP models and ethical data practices. By providing real-time cyberbullying detection, this system helps create a safer digital environment.

KEYWORDS: Cyberbullying, Social Media, BERT, NLP, Semi-supervised learning , Twitter API.

I. INTRODUCTION

This study focuses on developing a machine learning model for the detection of cyberbullying content on social media. The process begins with data collection from platforms like Twitter, Facebook, and Instagram, where instances of cyberbullying are prevalent. The collected data undergoes preprocessing steps such as tokenization, stemming, and lemmatization to clean and standardize the text.

The significance of this research lies in its potential to enhance the detection and prevention of cyberbullying on social media platforms. By leveraging machine learning techniques, automated systems that can effectively **identify and mitigate cyberbullying, ultimately creating a safer online environment for users.**

OBJECTIVES:

Cyberbullying refers to the use of digital platforms to harass, threaten, or humiliate individuals. It can take various forms, such as hate speech, personal attacks, or spreading false information. Detecting such behavior is crucial to ensure online safety and mental well-being.

Develop a Robust Cyberbullying Detection Model: Implement various machine learning algorithms, including ensemble methods and deep learning models, to analyze text data and identify patterns indicative of cyberbullying. Optimize the model through hyperparameter tuning and validation to ensure robustness and generalization.

Enhance Data Preprocessing Techniques: Utilize noise reduction techniques to eliminate irrelevant information and apply semantic analysis to preserve the true meaning of the text. This includes tokenization, stemming, lemmatization(are text preprocessing techniques used in natural language processing (NLP) to reduce words to their base or root form), and normalization to prepare the data for effective analysis.

Address Data Imbalance: Apply techniques like oversampling, under sampling, and synthetic data generation (e.g., SMOTE) to address data imbalance. This will prevent the model from being biased towards the majority class and enhance its ability to detect rare instances of cyberbullying.

Develop Real-Time Processing Capabilities: Integrate real-time processing using streaming data platforms such as Apache Kafka.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304080

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

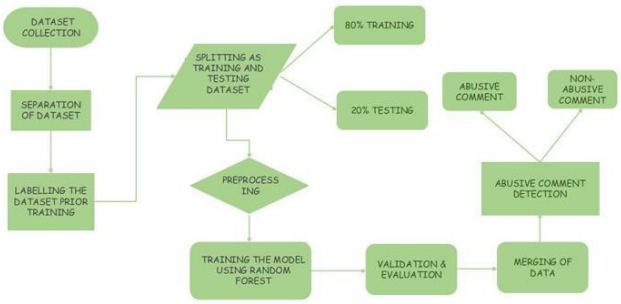
II. LITERATURE SURVEY

Title: Artificial Intelligence-based Cyberbullying Detection System on Social Media based on Machine Learning
Social media platforms have become integral to human communication and self-expression. However, the rise in connectivity has also led to an increase in harmful online behaviors, such as cyberbullying. With an abundance of user-generated content, detecting and mitigating abusive behaviors pose challenges. This literature review examines the application of prediction approaches, particularly utilizing deep learning technology, to assist in detecting cyberbullying. By analyzing textual data, user interactions, and contextual factors, advanced AI systems aim to protect individuals from online abuse while fostering safer digital environments.

Title: Advanced Cyberbullying Detection System Using Machine Learning.
The rapid growth of social media has amplified the importance of tackling cyberbullying to ensure user safety and mental well-being. However, identifying bullying content across diverse platforms and languages is a considerable challenge. This literature review explores a novel system that employs advanced deep learning techniques, including Convolutional Neural Networks (CNNs), to detect cyberbullying. By leveraging textual, visual, and user interaction data, the system provides accurate classification and flags harmful content. Such solutions empower social media platforms to create safer, more inclusive spaces.

Title: Automated Cyberbullying Detection Using Machine Learning
Automated cyberbullying detection using revolutionizes the identification of abusive behavior on social media platforms, process textual and visual content to classify cyberbullying efficiently, reducing reliance on manual moderation and enabling faster detection. This AI-driven approach enhances accessibility, particularly for smaller platforms with limited resources. By analyzing intricate patterns in online interactions, improve detection accuracy and help prevent the escalation of abuse. Furthermore, integrating models with real-time content monitoring systems allows for dynamic analysis, making cyberbullying detection more efficient, cost-effective, and impactful in fostering safe online environments.

ARCHITECTURE:



III. METHODOLOGY

System Development Approach :
The system leverages machine learning (ML) and deep learning (DL) techniques to enable efficient detection of cyberbullying across social media platforms. Python serves as the primary development environment due to its extensive ML/DL libraries, such as TensorFlow, PyTorch, and scikit-learn. Natural Language Processing (NLP) techniques are employed for analyzing textual data, while pre-trained DL models are used for feature extraction and classification. Social media platforms are integrated as primary data sources, ensuring real-time analysis of user interactions. The core functionality is implemented using Python scripting, enabling efficient input processing, text analysis, and classification tasks.

Data Input & Processing :
The system processes multiple types of inputs, including text, images, and metadata, to detect cyberbullying behavior:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Both types of data are processed to extract meaningful features. If bullying content is detected, the system flags it for review and prompts further action. Otherwise, it logs the analysis outcome as non-bullying content.

Detection and Classification Algorithms

The system employs both ML and DL techniques for cyberbullying.detection:

Natural Language Processing (NLP): Algorithms such as sentiment analysis and semantic analysis are used to analyze the tone and intent behind user-generated text.

Machine Learning Models: Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) provide high accuracy in contextual text analysis.

Ensemble Learning: Techniques like Random Forests or Gradient Boosting combine the outputs of multiple models to improve classification accuracy.

IV. SYSTEM ARCHITECTURE

The architecture consists of several key layers for efficient cyberbullying detection:

User Interface Layer: Provides web and mobile applications where users can report bullying or view flagged content.

Data Acquisition Layer: Continuously collects user interactions, comments, and posts from social media platforms.

V. CONCLUSION

In summary, this research highlights the promise of machine learning for detecting cyberbullying on social media, addressing challenges like data imbalance, contextual understanding, real-time processing, feature extraction, privacy concerns, and adversarial attacks. By leveraging advanced techniques, such as sentiment analysis, ensemble methods, deep learning models, and real-time processing, the study aims to enhance the accuracy and robustness of cyberbullying detection. The ultimate goal is to create automated systems that effectively identify and mitigate cyberbullying, fostering a safer online environment for users.Ethical handling of social media data is crucial to address privacy concerns.

Ultimately, this research aims to develop robust machine learning models that effectively identify and mitigate cyberbullying, creating a safer and more inclusive online environment for users. These advancements contribute significantly to the development of automated systems that enhance cyberbullying detection and prevention on social media platforms.

REFERENCES

1. Al-Harigy LM, Al-Nuaim HA, Moradpoor N, Tan Z (**2022)** Building towards Automated Cyberbullying Detection: A Comparative Analysis. *Computational Intelligence and Neuroscience, 2022*
2. Alom Z, Carminati B, Ferrari E (**2020**) A deep learning model for Twitter spam detection. Online Social Networks and Media 18:100079
3. Balakrishnan V, Khan S, Fernandez T, Arabnia HR (**2019)** *Cyberbullying detection on twitter using Big Five and Dark Triad features.* Pers Individ Differ 141, 252–257.
4. Bretschneider U, Wöhner T, Peters R (2014) Detecting online harassment in social networks.
5. I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "*Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio- Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.*
6. P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.
7. A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details