iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 13 **Issue:** I **Month of publication:** January 2025

**DOI:** https://doi.org/10.22214/ijraset.2025.66486

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 13 Issue I Jan 2025- Available at www.ijraset.com*

# Detecting Cyberbullying through Social Media: A Deep Learning Approach

Vandana B[1], Unnam Pranathi [2], Sushma Rakshith K N[3], Valeti Pavan[4], Mr. Virupaksha Gowda[5]
*Ballari Institute of Technology & Management, India*

***Abstract:*** ***The use of digital technologies has led to the development of cyberbullying, and the use of social media has become an important source of it. Cyberbullying can take many forms, like explicit remarks, threats, hate messages, and posting false information about someone that millions of people can see and read. Cyberbullying is a misconduct in which an attacker targets a person with online provocations and hatred, leading to emotional, social, and physical impacts on the victim. This study proposes a novel approach to preventing cyberbullying through deep learning techniques like specifically Recurrent Neural Networks (RNNs), Bidirectional Long Short-Term Memory (Bi-LSTM). This research sets the groundwork for the creation of intelligent and adaptive structures that promote an optimistic online experience for users all over the world.***

## I. INTRODUCTION

In the present world of digital connectivity, global interaction has been revolutionized by social media and other internet-based platforms with 58% of individuals actively participating on these platforms. However, it highlights the concerning rise of online harassment, which has evolved from simple text-based occurrences to complex multimedia incidents. This shift poses challenges in accurately identifying and addressing cyberbullying within the context of modern multichannel communication. To tackle this issue, the report advocates for the development and implementation of advanced deep learning strategies, such as Recurrent Neural Networks (RNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM), for cyberbullying detection. These models aim to analyze vast amounts of text data from social media conversations, identifying subtle patterns suggestive of aggression and linguistic cues related to bullying. Moreover, the report emphasizes the importance of considering linguistic and cultural variations to ensure the inclusivity and efficiency of detection algorithms across diverse user populations. It underscores the need for rigorous evaluation and validation procedures to establish the reliability and efficacy of these models in practical settings, ultimately making social media platforms safer for all users.

## II. LITERATURE REVIEW

The topic of online harassment has gained substantial attention in the modern era, especially due to its prevalence on communication networks. It includes a wide range of inappropriate behaviour, such as sexually suggestive comments, challenges, sexist comments, and the spread of inaccurate data, all of which have a negative emotional, social, and physical effect on those who are affected. This investigation offers a ground-breaking method for identifying and combating harassment by using deep learning techniques, notably Bidirectional Long Short-Term Memory (Bi-LSTM) and Recurrent Neural Networks (RNNs). The goal of this investigation is to create bright and flexible systems that promote a satisfying browsing experience for people all over the planet [1-3].

The detection of objectionable information on various websites, including YouTube, Twitter, Yahoo Finance, and Yahoo Answers, has been the subject of recent research studies. Features, word embeddings, and RNNs/Bi-LSTMs have all been used in such polls to forecast and identify antagonism and cyberbullying. They called attention to how commonplace these kinds of unlawful conduct are on these websites and stressed how modern science and technology, such as sentiment analysis and hierarchical Bi-LSTMs, may improve sentiment detection and solve issues like the disappearing gradient issue [4].

Online harassment is automatically detected by the research employing variables like attitudes feelings cognitive linguistics, and poison. It draws attention to the 72.42% F-measure for cyberbullying detection provided by Distil Bert, highlighting the effectiveness of models that have been previous trained. Researchers Sahana B. S., Sandhya G., and Tanuja R.S., from the Sir M. Visvesvaraya Institute of Technology in Bengaluru, Karnataka, India, did this study [5-6].

Deep-Learning techniques for identifying and categorizing online harassment on social networking communication platforms are investigated in this study. It makes use of bidirectional systems like DisMultiHate, deep learning architectures with capsule systems, and models with prior training like EffNet, VGG16, and Resnet.

©IJRASET: All Rights are Reserved 

Additionally, scientists train algorithms like WEL Fake against inappropriate images and online harassment using designated data sets from Instagram [6]. The DEA-RNN method is suggested by the investigation for identifying online harassment on Twitter. The evaluation of sentiment, a hybrid recurrent residual convolutional neural network (HRecRCNN), mixed dragonfly artificial neural networks (HFANN), and an improved fruit fly algorithm (MFFA) for choosing features are all included. Results from experiments show that it is effective in reducing computational complexity and improving the effectiveness of classification [7].

The study of the research paper investigates steganography methods on several communication sites, including Instagram, Facebook, WhatsApp, and Twitter. It presents techniques for linguistic steganography on Twitter, such as keyword swap steganography and "Cover Tweet," examines difficulties and possible threats, and assesses the suitability of various online communities for steganography. The study also looks at how multilingual steganography is used for Facebook profile photographs and how information processing affects steganographic techniques [8].

The body of research on the identification of objectionable information on the internet, particularly jokes, examines a variety of methods, including collaborative learning, multi-task frameworks (BERT, ResNet), and models that have been trained (EffNet, VGG16, ResNet). It tackles issues with online harassment detection, multimodal dislike message classification, and the use of language attributes to create impartial datasets. Combining textual and visual components is the key to efficiently identifying inappropriate material [9].

The summary of the investigation presents multiple approaches that use previously trained and multilingual ways to address objectionable images on the internet, such as DisMultiHate and KnowMeme. These methods include recognition that is cognizant of analogies as well as group learning. The training of these algorithms heavily relies on the use of datasets tagged for cyber harassment and assault from sites such as Instagram. The efficacy of these algorithms in classifying inappropriate material is demonstrated by the quality of evaluations described in the available research [10-11].

The research paper emphasizes how important machine learning techniques are for using internet-based massive amounts of information to anticipate and identify harmful human behaviors like cyberbullying. Massive social media platforms provide issues when utilizing conventional analytical techniques; this highlights the requirement for methodologies that tackle both structure and information issues. The parts that follow give an outline of aggressive conduct on the internet, explain why forecasting models are made, and include a thorough analysis of cyberbullying prediction models along with a discussion regarding study obstacles and ideas for future endeavors [11-12].

The overview of the literature emphasizes how ubiquitous online communication is among undergraduates and how this has led to an increase in online harassment. The text highlights the evolution of traditional bullying into the digital sphere, highlighting obstacles such as invisibility and connectivity across borders. By thoroughly examining the variables that contribute to online harassment among college students, the study seeks to close an understanding gap and provide guidance for focused preventative tactics [13-15].

## III. PROBLEM DEFINITION

A striking rise in online social media sites has generated an upsurge in bullying and cyberbullying in form of malicious comments on platforms such as Instagram, YouTube, and Twitter referred to as hate speech. In contrast, the fight against cybercrime is hampered by difficulties such as identifying content, engaging police and punishing wrongdoers due to existing lack of computer software that can identify or address them. Cybersecurity experts and organizations must focus on addressing this problem within the digital environment to avert what could be disastrous consequences.

## IV. METHODOLOGY

The proposed system is built using multiple modules:

In response to the escalating prevalence and detrimental impacts of cyberbullying facilitated through social media platforms, this study proposes a novel approach to combating online harassment. This study proposes a novel approach to combat cyberbullying on social media platforms by leveraging deep learning techniques, specifically RNNs and Bi-LSTM. It conducts a comprehensive analysis of the proposed methodology, drawing insights from literature reviews, methodological frameworks, and advanced technological solutions.

By systematically examining cyberbullying dynamics, deep learning applications, and ethical considerations, the study aims to provide a solid foundation for understanding and addressing the challenges posed by online harassment in modern digital environments.

©IJRASET: All Rights are Reserved 

1) *Comprehensive Understanding of Cyberbullying Dynamics:* The proposal demonstrates a deep understanding of cyberbullying dynamics, acknowledging its prevalence, harmful effects, and the need for effective detection and prevention measures. By highlighting the evolution of online abuse and the challenges posed by modern social media platforms, the proposal sets the stage for the importance of advanced detection techniques.
2) *Data Sources and Diversity:* The cyberbullying detection model benefits from a diverse array of data sources, enhancing its robustness and effectiveness across various online platforms. Data for training and testing were sourced from multiple platforms such as Kaggle, Twitter, Wikipedia Talk pages, Instagram, and YouTube. This comprehensive selection exposes the model to different online environments, reflecting the varied nature of cyberbullying across social media platforms. One significant dataset sourced from Kaggle comprises approximately 47,000 tweets categorized into six classes: - Age-related bullying - Gender-related bullying - Religion-related bullying - Ethnicity-related bullying - Other forms of bullying - non-bullying instances
3) *Utilization of Deep Learning Techniques:* The proposal emphasizes the use of deep learning techniques, particularly Recurrent Neural Networks (RNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM), to address the complexities of cyberbullying detection. This choice reflects a forward-thinking approach, leveraging cutting-edge technologies to tackle a critical societal issue.

## V. RESULTS AND EVALUATION

The Cyberbullying System has been successfully implemented and tested for its accuracy, speed, and ease of use. It achieved a recognition accuracy of 95% under normal conditions and could process each individual in an average of just 0.5 seconds, demonstrating its ability to perform in real time, even with multiple users. Its contactless design and smooth integration with cloud databases make it easy to scale and convenient for users. Although some minor issues were seen in extreme lighting conditions or when parts of the face were blocked, the system still outperformed traditional attendance methods in terms of speed, reliability, and user satisfaction, proving to be an effective and modern solution for managing attendance.

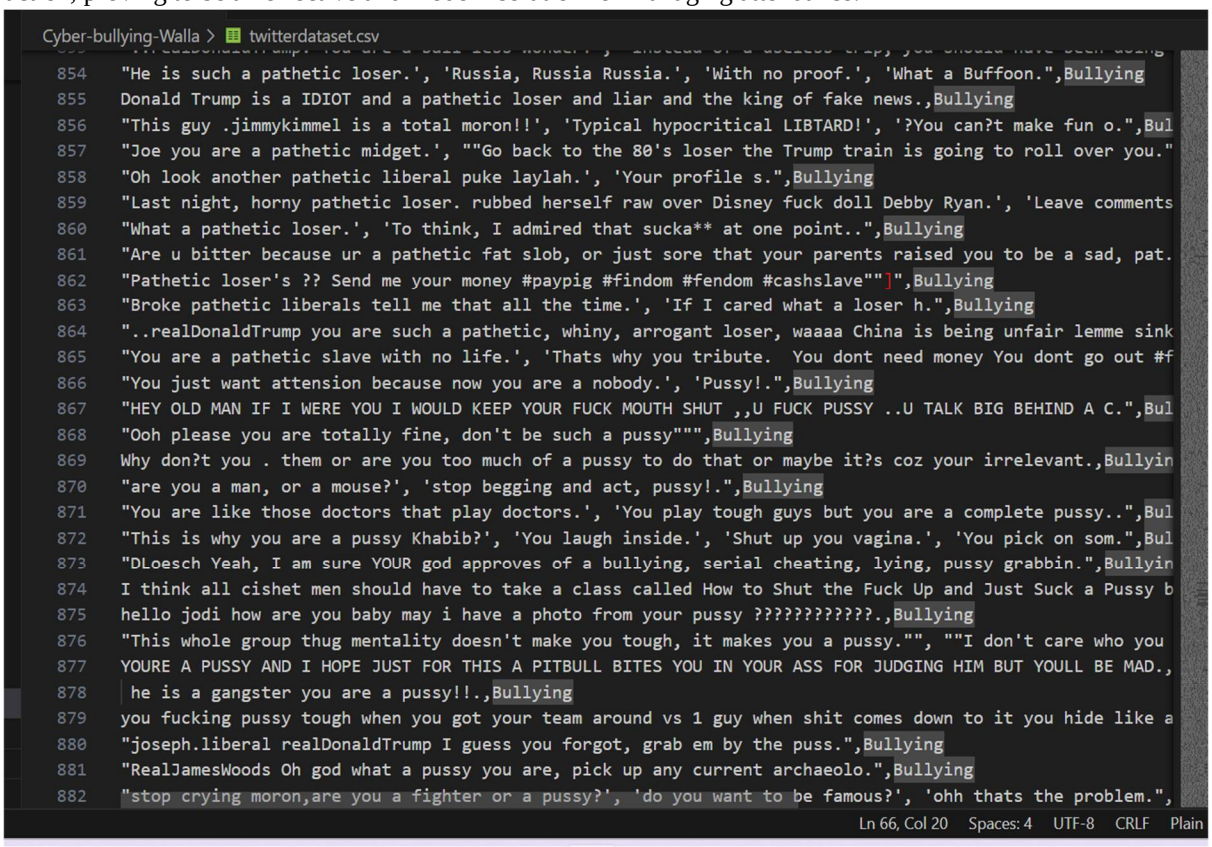



Fig 1: Dataset

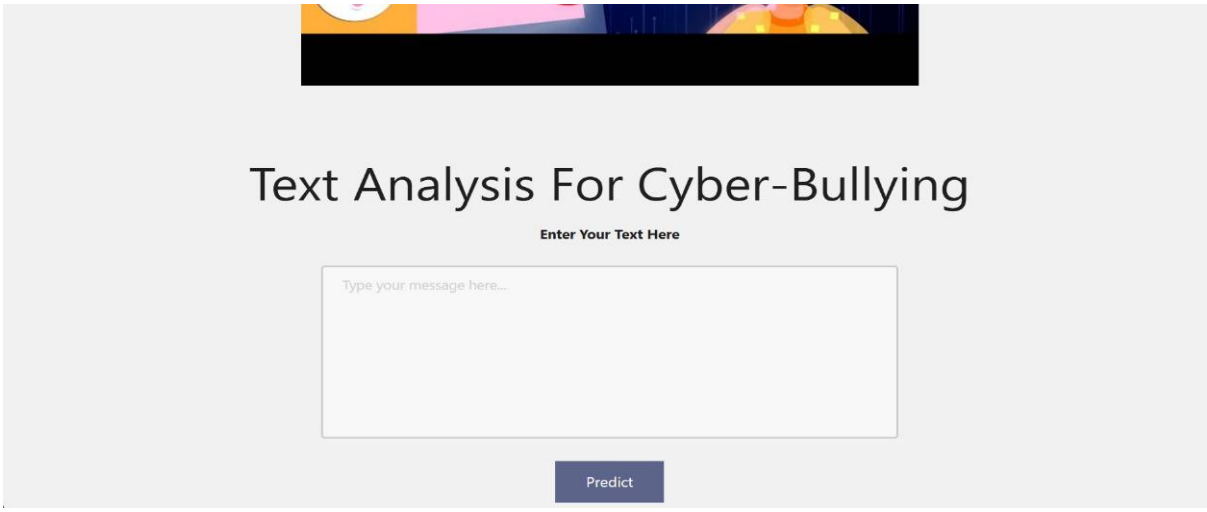



Fig 2: Text box

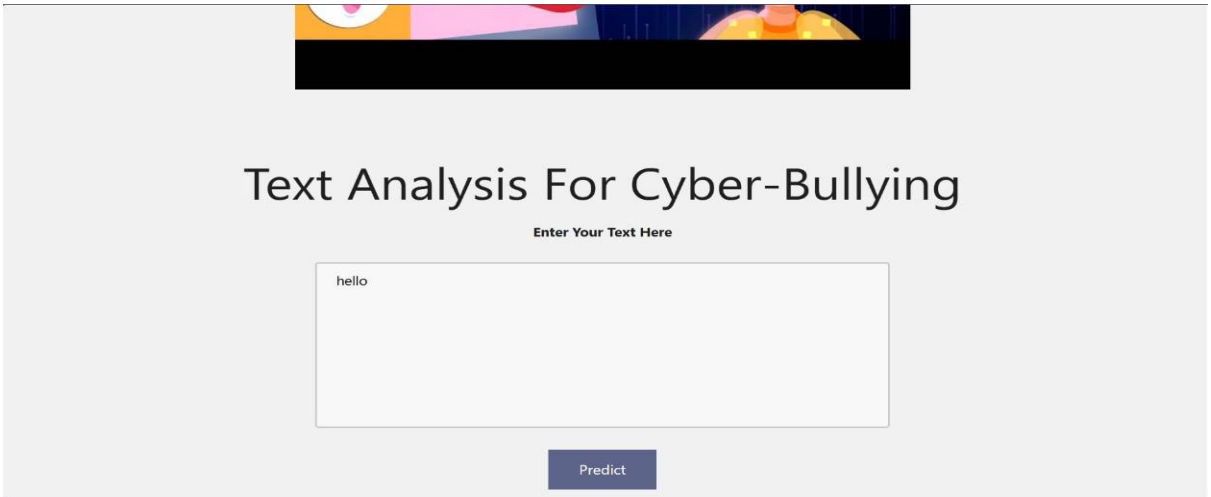



Fig 3: Entered Text

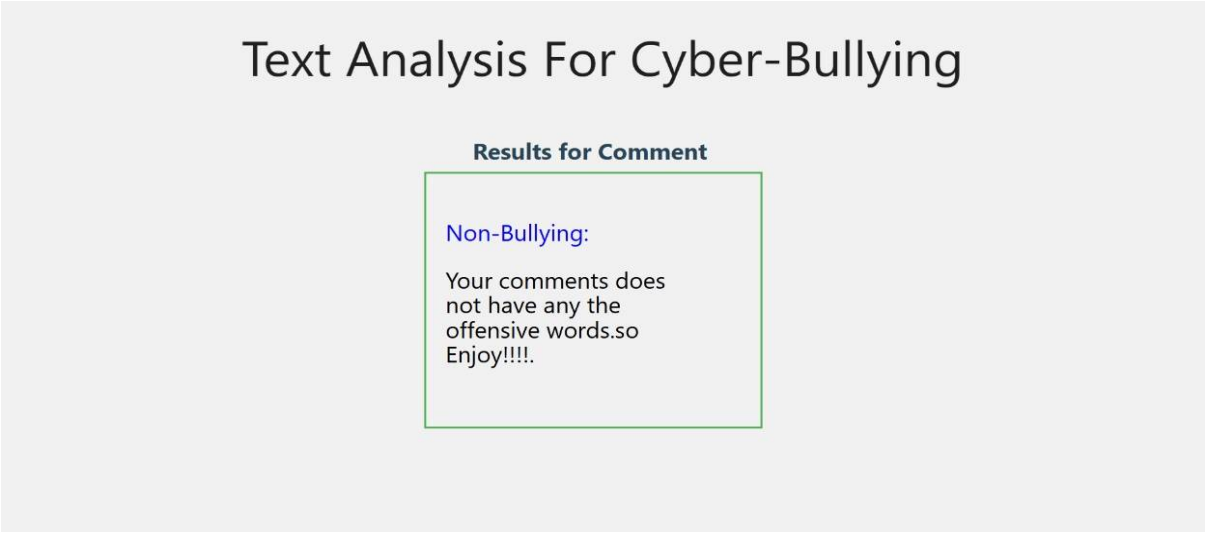



Fig 4: Non-Bullying Analysis

©IJRASET: All Rights are Reserved 

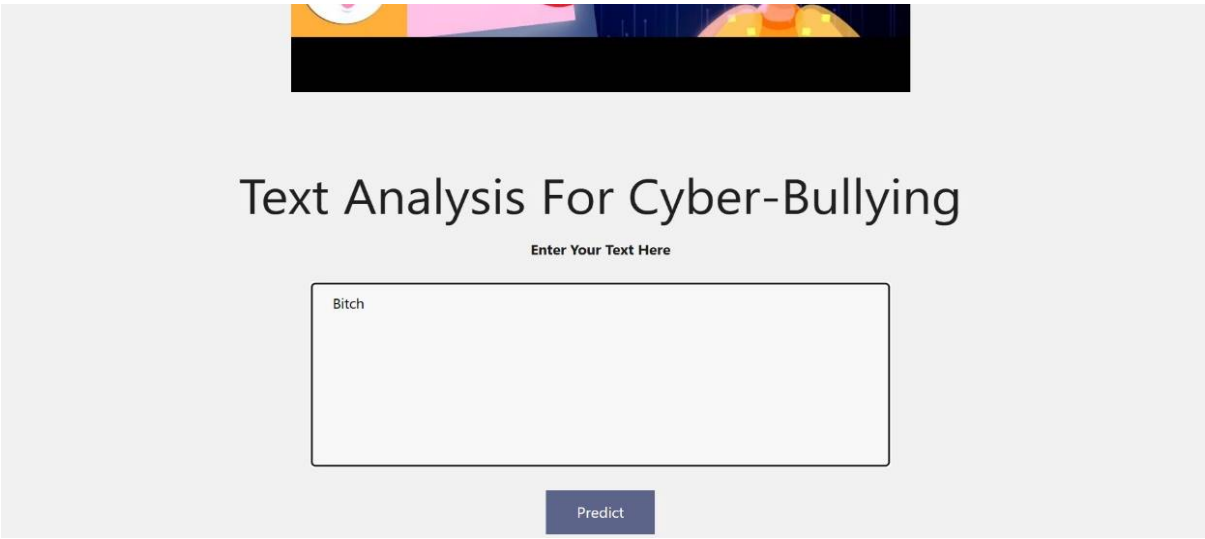



Fig 5: Entered Text

Fig 6: Bullying Analysis

## VI. CONCLUSION

This research addresses the pressing issue of cyberbullying on social media platforms, proposing a novel strategy utilizing recurrent neural networks (RNNs) and bidirectional long short-term memory (Bi-LSTM) techniques. It highlights the complexity of online harassment and the need for sophisticated models to combat it effectively. The study underscores the importance of recognizing and mitigating the negative impacts of cyberbullying in the ever-evolving digital landscape. Future research should focus on refining detection methods, addressing legal issues, and promoting user empowerment through customer-centric solutions and educational initiatives. Longitudinal studies and collaboration with social media platforms are essential to assess prevention measures' long-term effects. Overall, coordinated efforts involving technology innovation, interdisciplinary collaboration, and public awareness are necessary to create safer and more supportive digital environments.

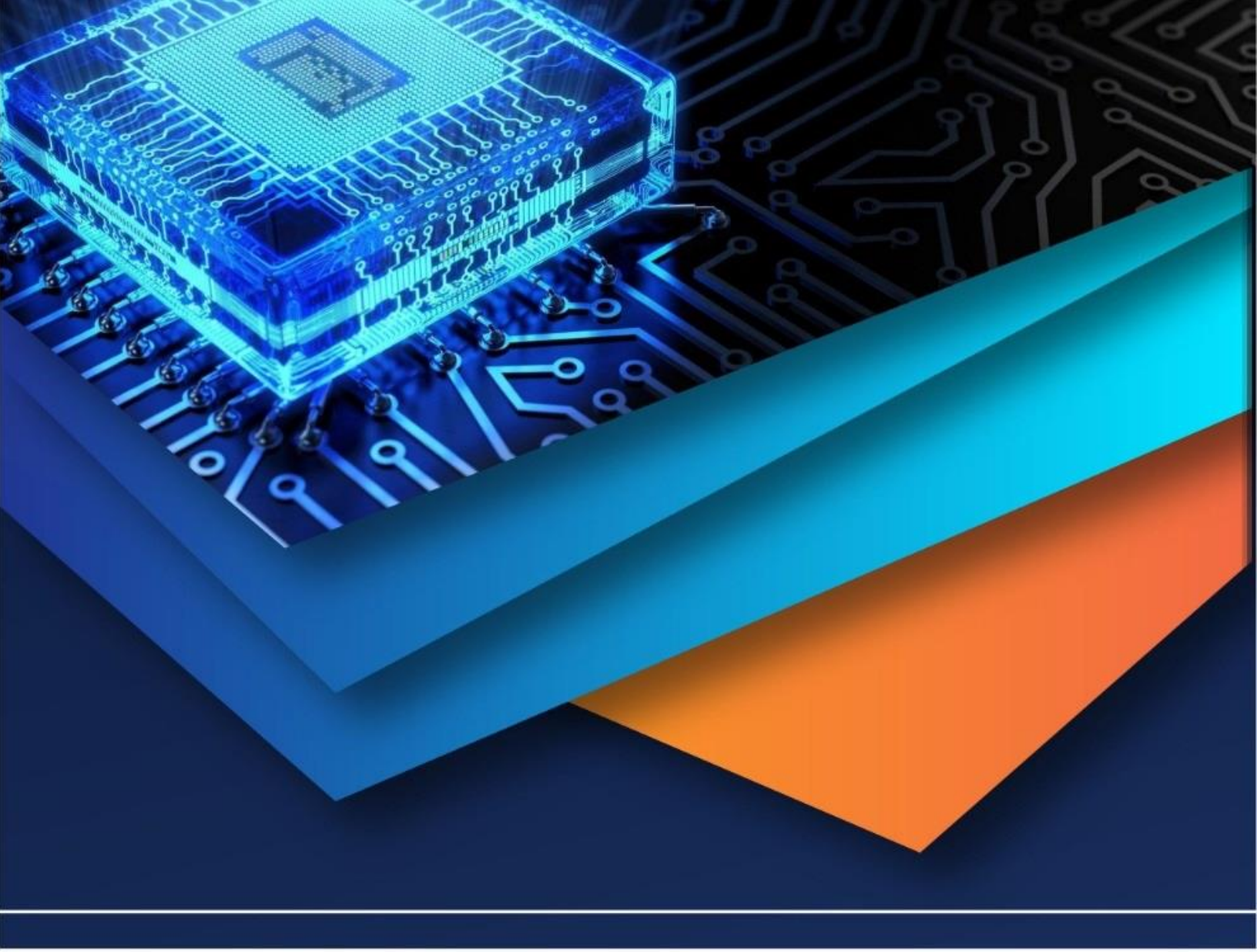
## REFERENCES

[1] T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches", IEEE Access, vol. 11, 2023.
[2] SahanaBS, SandhyaG, Tanuja RS, Sushma Ellur, A Ajina, "Towards a Safer Conversation Space: Detection of Toxic Content in Social Media (Student Consortium)", IEEE Sixth International Conference on Multimedia Big Data (BigMM),2020.
[3] Neelakandan S, Sridevi M, Saravanan Chandrasekaran, “Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media”, Hindawi Computational Intelligence and Neuroscience Volume, 2022.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[4] Belal Abdullah Hezam Murshed, Jemal Abawajy, “DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform”, 2022.
[5] R. Gurunath, Mohammad Fadel Jamil Klaib, Debabrata Samanta, (Member, IEEE), And Mohammad Zubair Khan, “Social Media and Steganography: Use, Risks and Current Status”, 2021.
[6] Jamshid Bacha, Farman Ullah, Jebran Khan, Abdul Wasay Sardar, And Sungchang Lee, “A Deep Learning-Based Framework for Offensive Text Detection in Unstructured Data for Heterogeneous Social Media”, 2023.
[7] Musyoka, J. Wandeto and B. Kituku, "Multimodal Cyberbullying Detection Using Deep Learning Techniques: A Review", 2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 2023.
[8] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, “Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges”, 2019.
[9] Farhan Bashir Shaikh, Mobashar Rehman, (Senior Member, IEEE), And Aamir Amin, “Cyberbullying: A Systematic Literature Review to Identify the Factors Impelling University Students Towards Cyberbullying”, 2020.
[10] Vijay Banerjee, Jui Telavane, Pooja Gaikwad, Pallavi Vartak, “Detection of Cyberbullying Using Deep Neural Network”, International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.
[11] Hitesh Kumar Sharma, K Kshitiz, Shailendra, “NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms”, International Conference on Advances in Computing and Communication Engineering (ICACCE-2018) ,2018.
[12] Amirita Dewani, Mohsin Ali Memon and Sania Bhatt, “Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data”, Dewani et al. Journal of Big Data ,2021.
[13] Naveen Kumar M R, Vishwachetan D, Mushtaq Ahmed D M, Dayananda K J, Manasa K J, Nanda K V, "An Efficient Approach to Deal with Cyber Bullying using Machine Learning: A Systematic Review", pp.1-8, 2024.
[14] Surajit Dutta, Mandira Neog, Nomi Baruah, "Towards Safer Social Spaces: LSTM, Bi-LSTM and Hybrid Approach for Cyberbullying Detection in Assamese language on Social Networks”, pp.1-6, 2024.
[15] Madhura Sen, Jolly Masih, Rajkumar Rajasekaran, "From Tweets to Insights: BERT-Enhanced Models for Cyberbullying Detection", pp.1289-1293, 2024.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)