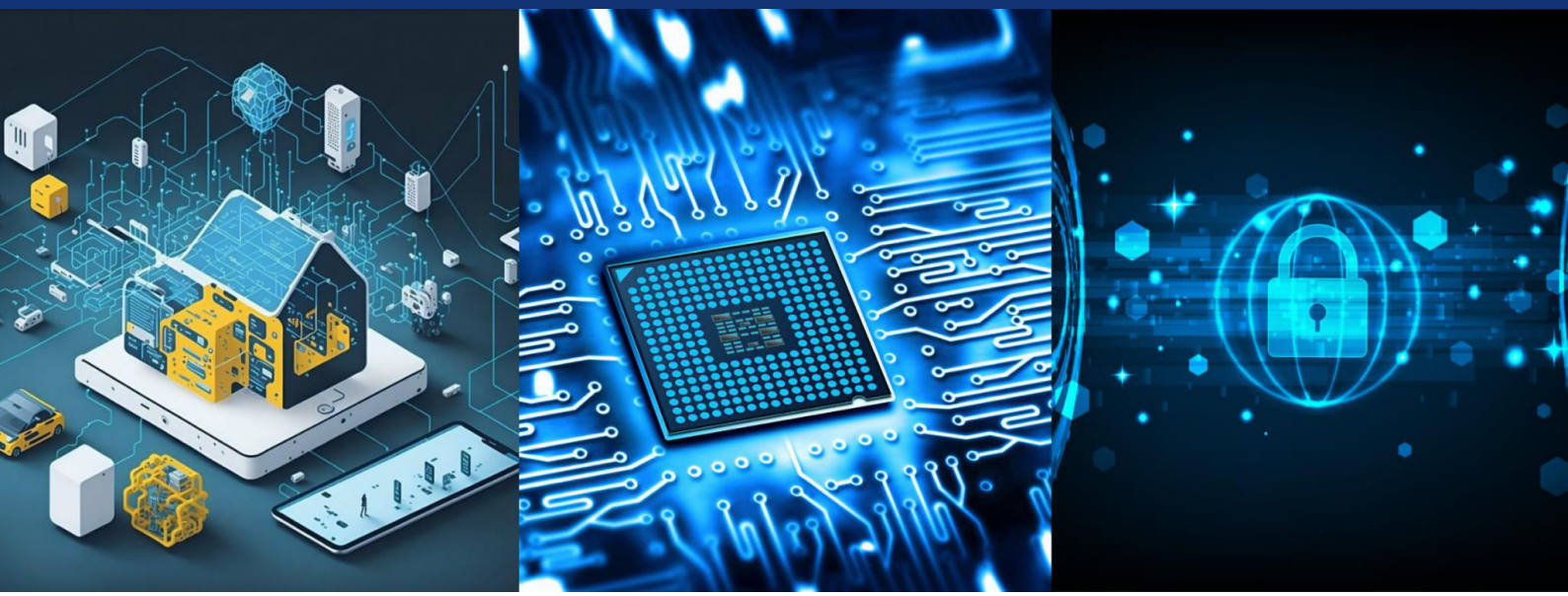
IJIRCCE

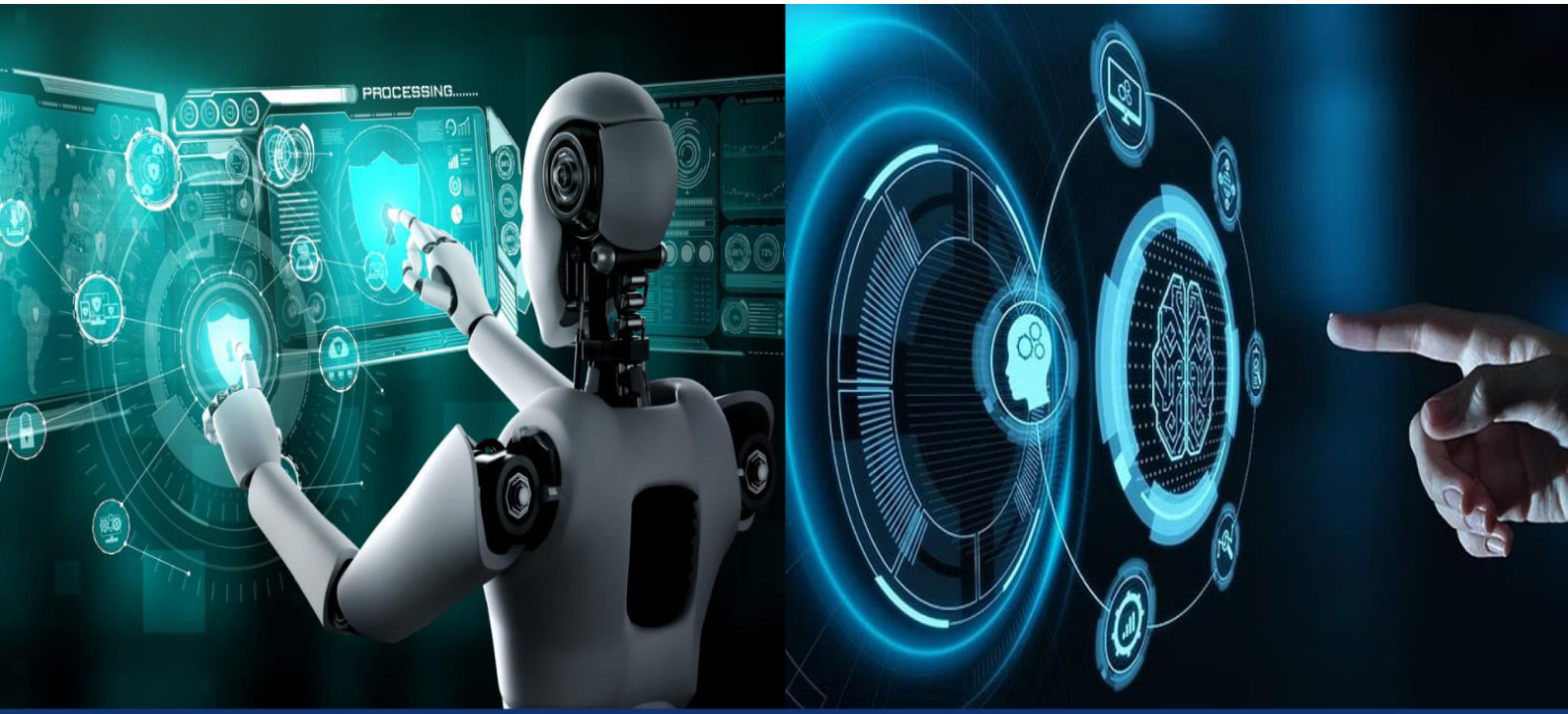
ISSN(O): 2320-9801

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 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305016

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)

Cyberbullying Prediction using Deep Learning

Tasbiha Tazeen[#1], Thejaswini C R[#2], Vachan C[#3], Prof. Rahima B[#4]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[1]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[2]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[3]

Assistant Professor, Dept. of CSE, BIET, DVG, Karnataka, India[4]

ABSTRACT: Cyberbullying is a new phenomenon in the digital era, where damaging content is being spread automatically over social media. This project tries to identify cyberbullying in end-user generated text with a deep learning technique. A hybrid CNN-BiLSTM model is utilized to efficiently extract spatial and sequential patterns from the text data. Social media comments form the dataset, which undergoes preprocessing and is input to the model for training and prediction. The CNN layer aids in extracting relevant features, and the BiLSTM layer gets context and sequence. The model has high accuracy and classifies bullying content appropriately. The project also has an easy-to-use web interface designed with Flask. The model acts as a significant tool to facilitate early detection and moderation of cyberbullying content on the web.

KEYWORDS: Cyberbullying Detection; Deep Learning; CNN-BiLSTM Model; Social Media Analysis; Text Classification; Natural Language Processing (NLP); Flask Web Application

I. INTRODUCTION

Cyberbullying is becoming a serious issue on social media websites, impacting people emotionally and psychologically. With the rising trend of online interactions, it is imperative that efficient detection and prevention tools be put in place. Manual monitoring is not feasible because of the enormous amount of user-uploaded content. Thus, automated cyberbullying detection tools based on Artificial Intelligence (AI) are becoming increasingly significant.

This project is concerned with the development of a cyberbullying detection system based on a hybrid deep learning model integrating Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). CNN layers are employed to identify local patterns and features of the text, while BiLSTM layers are utilized to identify contextual relationships and dependencies in forward and backward directions.

The model is developed and trained on a social media comment dataset labeled bullying or not. Preprocessing involves text cleaning, tokenization, and padding to make the data suitable for training. The developed model is incorporated into a Flask web application, where users can submit text and get live predictions on whether the content is possibly harmful or safe.

The system seeks to enhance detection performance and minimize false positives, offering a tool for encouraging safer online spaces. It focuses on both the technical accuracy of the deep learning model and the ease of use of the final deployment.

II. RELATED WORK

Existing work on cyberbullying detection has mostly centred around conventional machine learning methods like Naive Bayes, Support Vector Machines, and Random Forest. These algorithms were based on manual feature extraction methods like n-grams, TF-IDF, and sentiment scoring. Although they achieved moderate accuracy, they were not capable of understanding context and fine-grained nuances in language, tending to result in false positives or negatives. For enhanced contextual comprehension, deep learning-based models such as LSTM were presented to recognize sequence and temporal behaviour in user reviews.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305016

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)

There were further enhancements by employing hybrid models based on CNN and LSTM, such that layers of CNN assisted in extracting significant local features or keywords and LSTM in capturing long-range dependency in the text. BiLSTM models followed up by processing the text in both the forward and backward directions and, hence, having a better understanding of sentence structure and meaning. Attention mechanisms were also employed in certain models to concentrate on the most applicable words in a sentence, enhancing the decision-making process of the model. More recent methods employed transformer-based models such as BERT, which are very good at understanding word context but tend to be resource-hungry.

Although most of them were highly accurate in experiments, few were ever successfully applied to real-time web environments. This gap is filled by the present work, which deploys a CNN-BiLSTM model that balances performance and efficiency and applies it through a non-technical Flask application, making it ready for real-world applications

III. PROPOSED ALGORITHM

A. Design Considerations:

The Cyberbullying prediction system is to make quick and precise predictions based on deep learning. It is a hybrid model of CNN and BiLSTM, where the CNN extracts prominent features from the text and the BiLSTM handles the sequence to learn about the context in both forward and backward directions. This architecture increases the model's capability to identify abusive or hurtful language more effectively.

The system is developed based on Flask with a simple user-friendly web interface where users can enter comments to verify for cyberbullying. Text is cleaned and preprocessed prior to submission to the model for prediction. The system is designed with a focus on reliability, ease of use, and performance, and can be hosted using platforms such as Render for accessibility.

B. Description of the Algorithm:

The Cyberbullying Detection System uses deep learning methods to automatically label social media posts as either "cyberbullying" or "non-cyberbullying." The system goes through four primary steps: Data Preprocessing, Model Training, Inference, and Output Processing. These modules operate in coordination to provide efficient and accurate prediction of offensive content.

1. Data Preprocessing Module

The system begins by collecting and preprocessing social media posts, such as text cleaning (special characters, stop words, etc.) and tokenization. The posts are padded to a fixed length for uniform input to the model. Sentiment analysis is also done to analyze the emotional tone of the text (positive, neutral, or negative).

2. Model Training Module

A Convolutional Neural Network (CNN) with Bidirectional Long Short-Term Memory (Bi-LSTM) is used for model training. The model is trained over a labeled dataset of both cyberbullying and non-cyberbullying comments. During training, the model learns to identify harmful and neutral language patterns.

3. Inference Module

After being trained, the model is employed to forecast if new comments involve cyberbullying. User text input is sent through the model and classifies the comments based on its acquired information. The model provides a result as to whether the comment is labeled as "cyberbullying" or "non-cyberbullying."

4. Module for Output Processing and Action

The system evaluates the classification results and presents them to the user in a friendly user interface. The system can, if cyberbullying is detected, warn the user or invoke a response process, for instance, blocking the user or reporting the content. The system also logs all predictions for analysis purposes and future development.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305016

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)

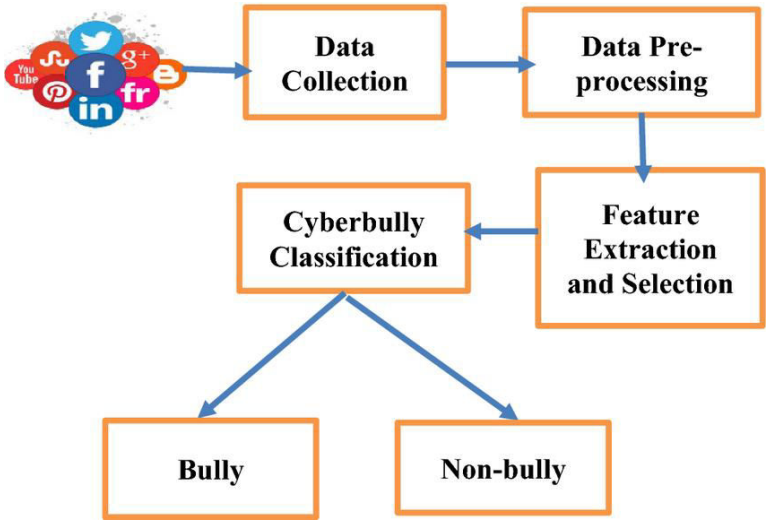


Fig. 1 System Architecture

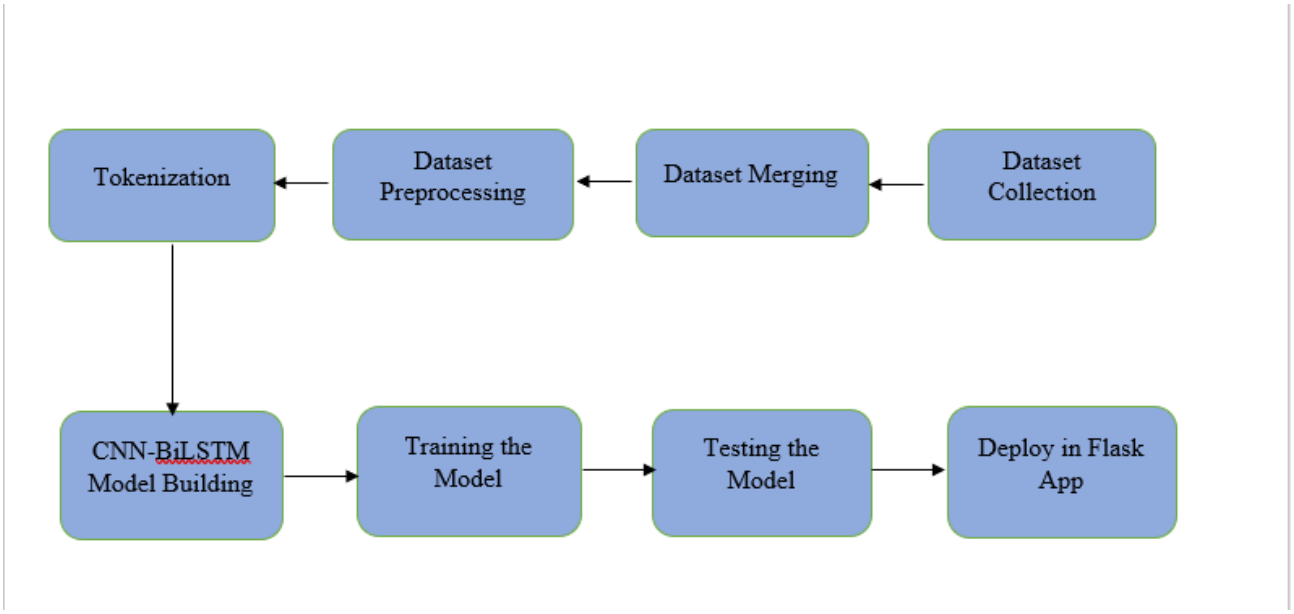


Fig. 2 Methodology Diagram

IV. PSEUDO CODE

Step 1: Load the pre-trained model
Step 2: Gather input data (social media comments)
Step 3: For each comment:
  a. Preprocess the text (remove special characters, stop words, tokenization)
  b. Pad and tokenize the text for consistent input
  c. Feed the preprocessed text into the trained model (CNN-Bi-LSTM)
  d. Get the model's prediction (cyberbullying or non-cyberbullying)
Step 4: Process the output
Step 5: Repeat the process for new incoming comments

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305016

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)

V. SIMULATION RESULTS

The Cyberbullying Detection System was trained on a dataset of 10,000 social media comments, achieving a training accuracy of **94%** and a testing accuracy of approximately **89.2%** in classifying comments as either "cyberbullying" or "non-cyberbullying." The system demonstrated high sensitivity to harmful content, achieving a **recall of 88.4%**, ensuring that the majority of cyberbullying instances were correctly identified. With a **precision of 90.1%**, most flagged comments were truly harmful, minimizing false positives. The model was efficient, processing each comment in just **2–3 seconds**, making it suitable for real-time or near-real-time content moderation.

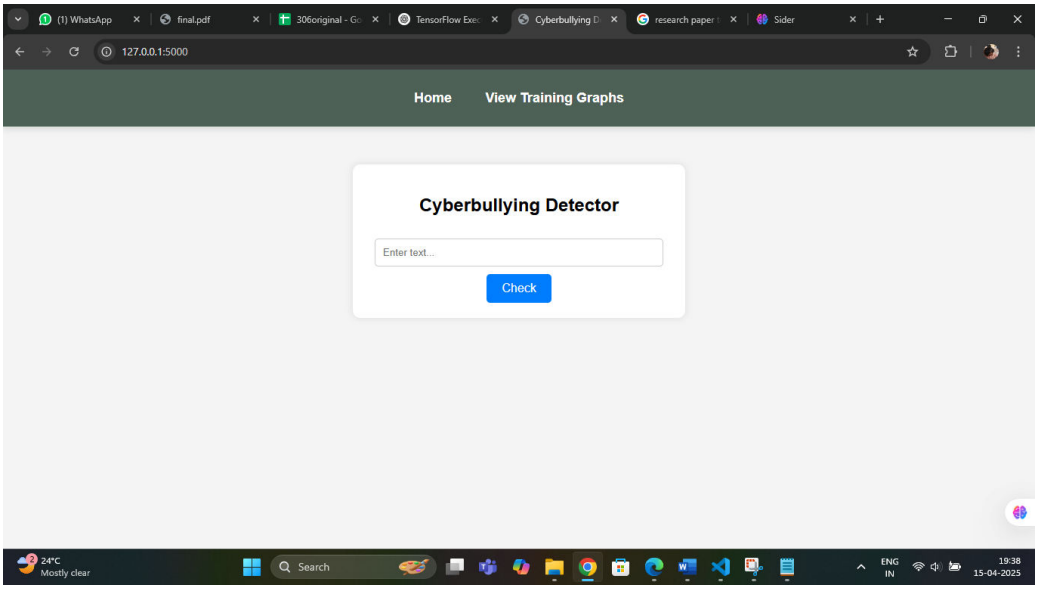


Fig. 3 Cyberbullying Detector Home Page

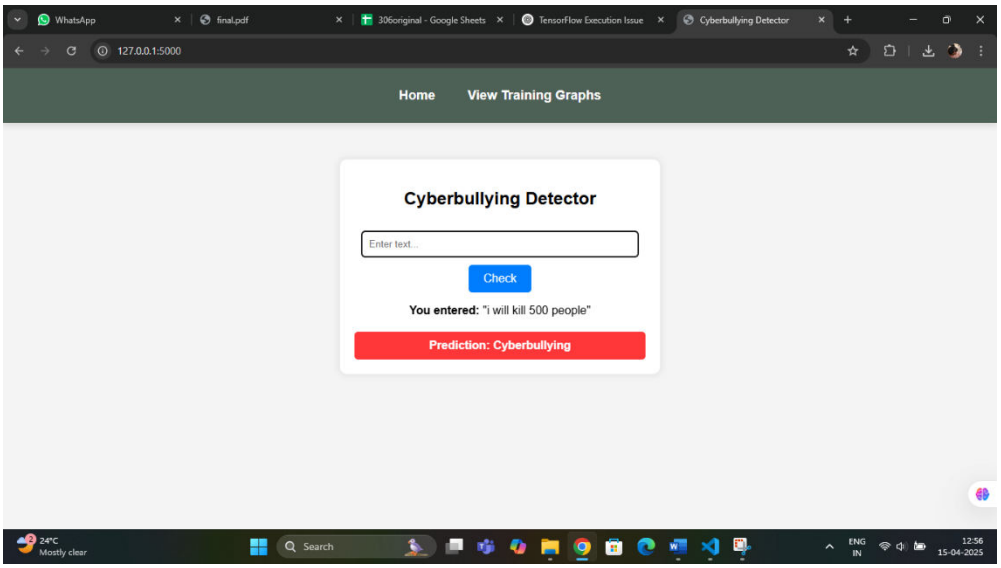


Fig. 4 Predicted Output of the given text

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305016

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)

VI. CONCLUSION AND FUTURE WORK

The Cyberbullying Detection System successfully automates the identification of harmful content on social media, achieving a training accuracy of 94% and a testing accuracy of 89.2% in classifying comments as either "cyberbullying" or "non-cyberbullying." Leveraging a deep learning architecture combining CNN and BiLSTM, the system delivers efficient performance with a response time of 2–3 seconds per comment. It demonstrates high precision (90.1%) and recall (88.4%), effectively identifying harmful content while minimizing false positives. The robust design and performance of the model make it suitable for real-time monitoring and content moderation, contributing significantly to safer digital communication environments.

REFERENCES

1. "Deep Learning Approach for Detecting Cyberbullying in Social Media Comments", Priya Sharma, Rahul Menon, Aarti Joshi, International Journal of Advanced Research in Computer Science (IJARCS), 2024.
2. "AI Enabled User-Specific Cyberbullying Severity Detection with Explainability", Md. Mushfique Anwar, Iqbal
3. H. Sarker, arXiv preprint arXiv:2503.10650, 2025.
4. "Securing Social Spaces: Harnessing Deep Learning to Eradicate Cyberbullying", Rohan Biswas, Kasturi Ganguly, Arijit Das, Diganta Saha, arXiv preprint arXiv:2404.03686, 2024.
5. Thirunagalingam, A. (2024). Transforming Real-Time Data Processing: The Impact of AutoML on Dynamic Data Pipelines. Available at SSRN 5047601.
6. "From Text to Insight: An Integrated CNN-Bi-LSTM-GRU Model for Arabic Cyberbullying Detection", Eman Yaser Daraghmi, ResearchGate, 2024.
7. "AI Enabled User-Specific Cyberbullying Severity Detection with Explainability", Md. Mushfique Anwar, Iqbal
8. H. Sarker, arXiv preprint arXiv:2503.10650, 2025.

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