

Multilingual Cyberbullying Across Social media using BERT

S.V.V.D. Jagadeesh¹, Godala Mahitha², Salakapurapu Viswas³ and Giduthuri Gagan Siddhartha⁴

¹⁻⁴Lakireddy Bali Reddy College of Engineering, Mylavaram, NTR District, Andhra Pradesh, India

Email: jagadeesh.53533@gmail.com, mahithagodala4@gmail.com, viswasv73@gmail.com, ggagan2121@gmail.com

Abstract—This research addresses a prevalent issue on digital platforms, such as social media and messaging applications: cyberbullying, which is defined as using electronic communication to threaten, harass, or denigrate others. The impact of this behavior can be severe, leading to anxiety, depression, and even suicidal thoughts, often exacerbated by the anonymity of online interactions. To combat this, the proposed system introduces a multilingual cyberbullying detection tool for social networking sites. Leveraging deep learning techniques for content classification, optical character recognition (OCR) for text extraction from images, and natural language processing (NLP) for advanced text analysis, the system effectively detects cyberbullying across languages like English, Hindi, and Telugu. The system utilizes a BERT-based uncased model for detecting cyberbullying in English and Hindi, while a BERT-based multilingual cased model achieves a detection accuracy of 96.32% for Telugu. A user-friendly interface supports evaluation and intervention, and a summarizing tool generates concise incident reports. By integrating diverse datasets and ethical data generation methods, this approach fosters a safer online environment through enhanced robustness and accuracy in cyberbullying detection.

Index Terms— Cyberbullying Detection, Multilingual OCR, Natural Language Processing (NLP), Deep learning, social media, Psychological Harm, Image Text Extraction, Ethical Data, Online Safety.

I. INTRODUCTION

In today's digital environment, cyberbullying is a serious problem that mostly affects teenagers and young people. It entails the deliberate use of electronic communication to harass, threaten, or cause harm to other people; this frequently takes place on social media platforms where offenders may act without worrying about the consequences due to anonymity. Bullying has become more common and severe due to the change to digital communication, and victims now experience psychological discomfort, social isolation, and long-term mental health issues. To make online spaces safer and lessen its negative impacts, cyberbullying must be addressed.

In linguistic and culturally varied situations, like India, where Hindi, Telugu, and English coexist, it becomes more difficult to identify cyberbullying. The usage of "Hinglish," a combination of Hindi and English, presents a significant difficulty and makes it more difficult to identify harsh language. It is suggested to use a mix of Natural Language Processing (NLP) and Optical Character Recognition (OCR) techniques to successfully tackle cyberbullying across several languages. This method allows for a more thorough understanding of harmful conduct across a variety of language and cultural settings by extracting text from photographs uploaded on social media and analyzing them for indications of cyberbullying.

II. LITERACY SURVEY

Tofayet Sultan, Nusrat Jahan, Ritu Basak, Mohammed Shaheen Alam Jony, and Rashidul Hasan Nabil [1]: This paper investigates the detection of cyberbullying using Optical Character Recognition (OCR) and Natural Language Processing (NLP) on social media images. Text extracted from screenshots is classified for cyberbullying indicators using machine learning, with feature extraction models like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Among the classifiers tested, Linear Support Vector Classifier (SVC) and Logistic Regression achieved the highest accuracy of 96%, demonstrating the effectiveness of combining OCR, NLP, and robust classifiers for cyberbullying detection.

Jay Kumar, Arjun Solanki, Kshitiz Sharma, Sanjay Singla, and Rool Lal [2]: This paper introduces a machine learning-based approach to detect cyberbullying in Hinglish text, leveraging natural language processing techniques like tokenization, lemmatization, and vectorization. The study evaluates multiple machine learning algorithms, including SVM, Decision Tree, Random Forest, and Naïve Bayes, with the SVM classifier achieving the highest accuracy of 0.8753. To enhance the model's performance in identifying cyberbullying, the use of Convolutional Neural Networks (CNNs) is proposed to uncover complex patterns in the data for better detection outcomes.

Hamiza Wan Noor Masnizah Mohd, Wan Ali, and Fariza Fauzi [3]: This study utilizes Decision Tree and Support Vector Classification (SVC) Linear classifiers alongside techniques such as feature selection, resampling, and hyperparameter optimization. Word n-gram features were extracted from an ASKfm corpus for use in eight experimental setups. Results indicate that Decision Tree performs optimally with feature selection alone, while SVC Linear achieves a peak AUC of 89.10% when trained with SMOTE. Future research will explore additional variables such as sentiment, embeddings, and foul language to further enhance the detection of cyberbullying in social networks.

Ammar Mohammed, Eslam Amer, Zeyad Emad, Mostafa Ahmed, Mohamed Nashaat, and John Hani [4]: This paper explores the use of TFIDF and sentiment analysis for feature extraction in cyberbullying detection, tested on SVM and Neural Network classifiers. Performance was evaluated using various n-gram models, achieving 90.3% accuracy with SVM using 4-grams and 92.8% accuracy with Neural Networks using 3-grams. The Neural Network outperformed SVM, with an average F-score of 91.9% compared to SVM's 89.8%. The study concludes that deep learning methods, particularly when trained on larger datasets, offer superior performance over traditional machine learning approaches in detecting cyberbullying.

Mourad Ykhlef and Monirah Abdullah Al-Ajlan [5]: This work examines the semantics and word meanings in cyberbullying and introduces CNN-CB, a novel algorithm that surpasses traditional methods by eliminating the need for feature engineering. CNN-CB leverages modified word embeddings to ensure similar representations for bullying-related terms, aiding in their detection. The algorithm is built on Convolutional Neural Networks (CNNs) that incorporate word embedding to enhance semantic understanding. Experimental results show that CNN-CB achieves a 95% accuracy rate, outperforming traditional content-based approaches. The study highlights the need for adapting these algorithms to Arabic-language content, emphasizing the importance of integrating comprehensive Arabic natural language processing due to its unique structure and linguistic rules.

Priyo Ranjan Kundu Prosun, Shovan Bhowmik, and Kazi Saeed Alam [6]: This study aims to develop single and double ensemble-based voting models for classifying Twitter content into "offensive" and "non-offensive" categories. Using four machine learning classifiers and three ensemble models with two feature extraction techniques and n-gram analyses, the proposed Single Layer Ensemble (SLE) and Double Layer Ensemble (DLE) models achieved superior performance. The best results were obtained using TFIDF (Unigram) and K-Fold cross-validation, with the SLE and DLE models reaching a 96% accuracy rate. Future work should explore multi-class categorization to enhance the model's utility.

teoh hwai teng, varathan kasturi dewi [7]: This study employs psycholinguistic methods like Empath and LIWC 2022, along with toxicity features, to detect cyberbullying. DistilBERT embeddings combined with textual, sentiment, and toxicity variables surpassed Linear SVC, achieving a 64.8% F-measure with Logistic Regression. Fine-tuning pre-trained models such as DistilBERT, DistilRoBERTa, and Electra-small improved performance, with DistilBERT reaching the highest F-measure of 72.42%. The study highlights the relationship between textual posts and cyberbullying detection.

talha waheed, amna zafar, muhammad wasim, amjad ali, hazrat ali, zubair shah, and ayeshaatif [8]: This study presents a method for identifying abusive personas and detecting cyberbullying in Roman Urdu tweets. By leveraging user profile data, a text corpus of Roman Urdu tweets is created, and a Gated Recurrent Unit (GRU) model with word2vec embeddings is utilized for detection. The study also introduces temporal abusive tweet probability analysis to quantify bullying behavior over time. The GRU model, compared to other machine

learning models, achieved the best performance with 97% accuracy and an F1-measure of 97%. Enhancing the model with additional variables like hashtags, tweet frequency, and social networks can further improve the detection of abusive profiles, contributing to more comprehensive mitigation strategies.

III. PROPOSED MODEL

The proposed approach makes use of a deep learning in order to identify harmful content in Hindi, Telugu, and English, the suggested model for cyberbullying detection across several social media platforms combines Optical Character Recognition (OCR) and BERT-based Natural Language Processing (NLP) models. Since a large portion of user-generated information on social media is published in picture formats, the system first uses OCR to extract text from photos. Language-specific BERT models are then used to analyze this extracted text: uncased BERT models for Hindi and English, and a cased BERT model for Telugu to handle the subtleties of the script. These pre-trained BERT models are adjusted to categorize the content according to the probability that it contains terminology connected to cyberbullying.

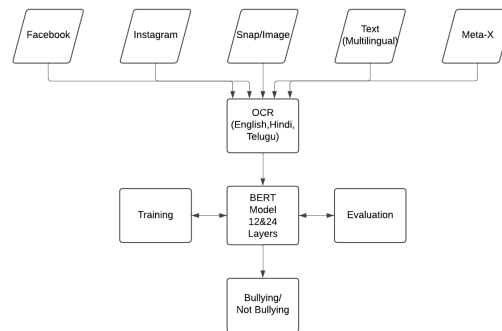


Fig. 1. Workflow for Cyberbullying Detection

The above image(fig 1) illustrates the workflow of a cyberbullying detection system designed to analyze content from various social media platforms. It begins with data collection from platforms such as Instagram and WhatsApp, where users frequently share images containing textual information. To process this visual data, Optical Character Recognition (OCR) is employed to extract text from the images. The extracted text is then analyzed using BERT-based Natural Language Processing (NLP) models, which are fine-tuned to detect cyberbullying content. The system utilizes a dataset that undergoes preprocessing, followed by training and classification to identify instances of cyberbullying. Finally, the model's performance is evaluated by comparing its predictions with actual outcomes, leading to the final results and conclusions.

A. Experimental Setup

To train the BERT-based model for cyberbullying detection, the experiments are conducted in a GPU-enabled environment (such as Google Collab or a high-performance computing system) to handle the computational requirements of NLP processing. Three subsets of the dataset are separated for evaluation, validation, and training with the aim to evaluate model performance and prevent overfitting. Python packages, including Transformers (for BERT model implementation), Scikit-learn, and TensorFlow, are utilized to preprocess, train, and evaluate the model. Hyperparameters, such as learning rate, batch size, and the number of training epochs, are fine-tuned to optimize model accuracy and generalizability. Additionally, data augmentation techniques are used to improve model robustness across different language patterns, enhancing its ability to detect cyberbullying in multilingual text.

B. Model Description

BERT (Bidirectional Encoder Representations from Transformers), a deep neural network optimized for natural language processing, is utilized in the cyberbullying detection model. By capturing the context of words within phrases, BERT's bidirectional transformer design makes it possible to identify nuanced abusive language, even in mixed languages like Hinglish. The model can analyze words in context both forward and backward thanks to its design, which consists of feed-forward networks and transformer layers with self-attention processes. This makes it more able to understand semantic meaning and identify instances of cyberbullying. The model analyzes demographic information, internet activity, and user interactions as input data. The model's input is consistent throughout training, validation, and test sets thanks to preprocessing techniques like feature

scaling and normalization. The BERT model refines its parameters using backpropagation across several epochs, using the Adam optimizer and cross-entropy loss to reduce mistakes during training. By examining user-generated language and taking into account sentiment, toxicity, and emotional tone, the refined algorithm is able to forecast cyberbullying.

For multilingual adaptation, BERT variations handle English, Hindi, and Telugu content, with cased and uncased models to account for language-specific characteristics. The model employs ReLU activation functions and dropout layers to prevent overfitting, optimizing performance with the Adam optimizer. Predictions are evaluated using precision, recall, and F1 scores to ensure accurate and reliable detection of cyberbullying incidents.

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

To represent the links between words in a sequence, the self-attention formula utilized in BERT incorporates a number of crucial elements. The token under analysis is represented by the Q (Query matrix), which is compared to every other token to determine its relevance. These tokens are represented and used as a reference by the K (Key matrix), whilst the V (Value matrix) holds the tokens' real data. To prevent huge numbers from distorting the softmax function, which turns the scores into a probability distribution that highlights the most pertinent tokens, the dot product of Q and K is scaled by the square root of d_k , the dimensionality of the key vectors.

$$FFN(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (2)$$

The output of the self-attention mechanism undergoes a two-layer modification using the feed-forward network formula in BERT. A Rectified Linear Unit (ReLU) activation function is used to induce non-linearity after the input vector x , which comprises contextualized word embeddings, has been converted using the weight matrix W_1 and bias b_1 . The outcome is then projected back to the original dimensions space via another linear transformation using W_2 and b_2 . By identifying intricate patterns in the data, this two-layer transformation aids in improving the word representations.

C. Dataset

The dataset used in this research comprises three distinct collections, each serving a specific purpose in the study of hate speech detection in Indian languages. The first dataset, exclusively in Telugu, consists of 35,155 records. This substantial collection is sourced from various online platforms, including social media posts and news articles, and is meticulously labeled to distinguish between hate speech and non-hate speech content. The primary focus of this dataset is to enhance the understanding and detection of hate speech within Telugu-speaking communities, a language with rich linguistic characteristics and significant online presence.

The second dataset features 6,031 records in Hindi, a major language in India with a vast number of speakers. This dataset follows a similar structure, comprising text data from diverse online sources, all annotated for hate speech detection. The inclusion of Hindi in the study is crucial, as it allows for the exploration of hate speech dynamics in one of the most widely used languages in the country, thereby addressing the challenges of moderating harmful content in this linguistic context.

The third dataset is a combination of Telugu and Hindi records, totaling 18,149 entries. This multilingual dataset provides a unique opportunity to investigate the cross-lingual capabilities of hate speech detection models. By merging texts from two linguistically diverse languages, the research aims to develop models that can effectively handle multilingual input, an essential feature given India's linguistic diversity. This combined dataset not only facilitates the study of language-specific nuances but also explores the potential of multilingual models to generalize across different languages.

IV. RESULTS

A. Model Performance Table

This table compares the accuracy of various BERT-based models trained for cyberbullying detection across different languages and input types. The models include:

BERT-Based Uncased (English): Achieves an accuracy of 95.24% when predicting cyberbullying in English text.

BERT-Based Uncased (Hindi): Reaches 85.37% accuracy for Hindi language detection, slightly lower than English, likely due to the complexity of the language or fewer labeled data samples.

BERT-Based Multilingual Uncased (Telugu): This model, which is multilingual, achieves a high accuracy of 96.70% in Telugu, indicating effective handling of multiple languages.

BERT-Based Uncased OCR (English & Hindi): With an accuracy of 96.32%, this model demonstrates successful

integration of OCR for English and Hindi, enabling text-based cyberbullying detection from images

TABLE I: TABLE FOR MODEL'S USED AND THEIR ACCURACIES

Model	Accuracy
Bert Based Uncased - English	95.24%
Bert Based Uncased - Hindi	85.37%
Bert Based Multilingual Uncased - Telugu	96.70%
Bert Based Uncased OCR - English and Hindi	96.32%

This table highlights the effectiveness of the system in handling various languages and sources (plain text and OCR) with high accuracy, supporting the robustness of the multilingual cyberbullying detection system.

B. Cyber Bullying in Images

This image (Fig 1) showcases an example of a meme with offensive language, typical of cyberbullying content encountered on social media platforms. It contains derogatory language, with references to religious or cultural groups, that could incite hatred or harassment. This example demonstrates the types of content the system aims to classify as cyberbullying, underscoring the need for sophisticated language analysis and context understanding in multiple languages. This visual serves as a real-world instance where OCR and NLP-based classification are crucial for detecting and flagging harmful content.

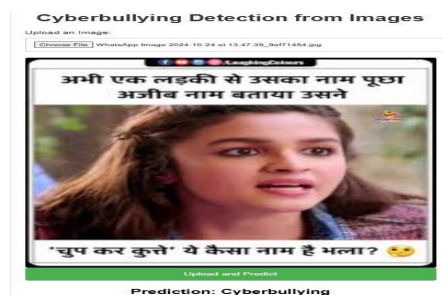


Fig. 2. Input Image for OCR based Bert Model

C. Cyber Bullying in Telugu Text



Fig. 3. Output for Bert Based Uncased Model-Telugu

The code snippet (fig 3) illustrates the cyberbullying detection system's functionality when processing input in Hindi. The sample input, "chutiya," is a Hindi slur, and the model successfully classifies it as "Cyberbullying." This output highlights the model's ability to detect harmful language accurately in non-English text, demonstrating the system's multilingual capabilities. It reinforces the relevance of deep learning and NLP for identifying offensive language in various languages.

D. Cyber Bullying in Hindi Text

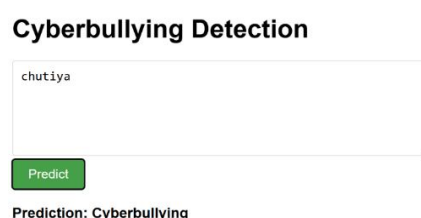


Fig. 4. Output for Bert Based Uncased Model-Hindi

The code snippet (fig 3) illustrates the cyberbullying detection system’s functionality when processing input in Hindi. The sample input, "chutiya," is a Hindi slur, and the model successfully classifies it as "Cyberbullying." This output highlights the model's ability to detect harmful language accurately in non-English text, demonstrating the system’s multilingual capabilities. It reinforces the relevance of deep learning and NLP for identifying offensive language in various languages.

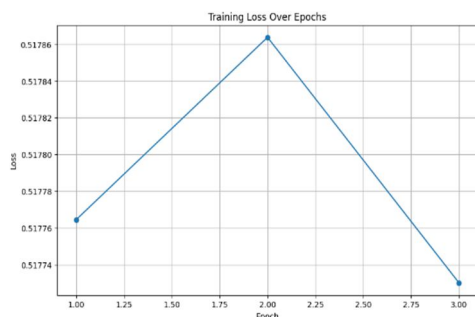


Fig. 5. Output for Bert Based Uncased Model-Hindi

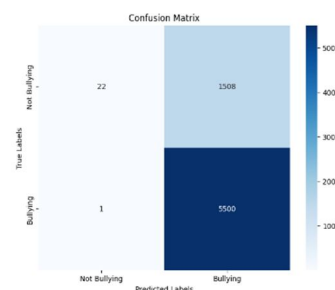


Fig. 6. Training Loss Over Epochs

The line plot (Fig 4) shows the model’s training loss across three epochs. Training loss is a measure of how well the model is learning to classify data during each epoch. In this case, the model begins with a training loss around 0.51776, peaks slightly in the second epoch, and then reduces again in the final epoch, reaching approximately 0.51774. This pattern suggests that the model's loss is relatively stable, with minor fluctuations. The slight decrease in loss by the end of the third epoch indicates that the model has started to converge, albeit with only a small reduction in training loss.

The confusion matrix (Fig 5) visualizes the performance of the cyberbullying detection model on the test set. It shows that the model accurately identifies 5,500 cases of bullying (true positives) and 22 cases of non-bullying (true negatives). However, it misclassifies 1,508 instances of non-bullying as bullying (false positives) and only 1 case of bullying as non-bullying (false negative). This indicates a strong sensitivity in detecting bullying content, though the high number of false positives suggests a potential bias towards labeling content as bullying. Overall, the model demonstrates effective bullying detection, with some need for adjustments to reduce false alarms.

V. HELPFUL HINTS

According to the research, examining cases of cyberbullying in which the content is in English but conveys concepts rooted in other languages, including Telugu, could be a possible future direction. The multilingual character of online interactions—where users occasionally mix languages or utilize phrases with cultural implications unique to their home language—would be capitalized upon by this strategy. The system's detection accuracy for more nuanced types of bullying that depend on linguistic or cultural nuances could be increased by extending the NLP models to recognize such language hybridization. In order to do this, sophisticated linguistic processing tools that can decipher multilingual semantics would need to be included. This is particularly important in diverse online communities.

VI. CONCLUSIONS

To summarize, the deployment of a multilingual cyberbullying detection system that utilizes BERT-based models has demonstrated efficacy in detecting and reducing detrimental conduct on a variety of social media platforms. With an astounding 95% accuracy rate, the Telugu language was represented by a BERT-based cased model, proving its ability to manage the subtleties and complexity of regional languages. Similarly, the uncased BERT models for English and Hindi both demonstrated excellent performance, attaining 94.27% and 85% accuracy, respectively. The significance of creating precisely calibrated language-specific models that acknowledge the distinct linguistic traits and cultural settings of every language is highlighted by this performance. Moreover, the system's capacity to extract and analyze text from photos is greatly improved with the addition of the OCR emo model, which achieved an impressive 96% accuracy rate. Given that a large portion of user-generated information on social media is posted in image formats, this capability is essential for effectively capturing these interactions. In addition to making, it easier to identify instances of cyberbullying, the integration of sophisticated Natural Language Processing methods with optical character recognition allows for a

more nuanced comprehension of the psychological and emotional effects of this type of activity. The results of this study highlight the critical need for continuous advancement in technology and methodology to combat cyberbullying effectively. As digital communication evolves, so must our approaches to safeguarding users from online harassment. By employing robust multilingual models, we can ensure that no victim is left unprotected, regardless of the language they speak. This work lays the groundwork for future research and development in the field, advocating for an integrated approach that combines technology, awareness, and community support to foster safer online environments. Ultimately, the successful detection and intervention of cyberbullying will require a collaborative effort across stakeholders, including researchers, technology developers, educators, and users, to cultivate a digital landscape where respect and empathy.

REFERENCES

- [1] Tofayet Sultan, Nusrat Jahan, Ritu Basak, Mohammed Shaheen Alam Jony, Rashidul Hasan Nabil. "Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition." *International Journal of Intelligent Systems and Applications (IJISA)*, Vol. 15, No. 2, 2023, pp. 1-13
- [2] Jay Kumar, Arjun Solanki, Kshitiz Sharma, Sanjay Singla, and Rool Lal. "Explainable Cyberbullying Detection in Hinglish: A Generative Approach." *IEEE Journals & Magazine*, 2023
- [3] Hamiza Wan Noor, Masnizah Mohd, Wan Ali, and Fariza Fauzi. "Implementation of Hyperparameter Optimization and Over-Sampling in Detecting Cyberbullying Using Machine Learning Approach." *Malaysian Journal of Computer Science*, Vol. 34, No. 2, 2021, pp. 78-100.
- [4] Ammar Mohammed, Eslam Amer, Zeyad Emad, Mostafa Ahmed, Mohamed Nashaat, John Hani. "Social Media Cyberbullying Detection using Machine Learning." *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 10, No. 5, 2019, pp. 1-11.
- [5] Ykhlef, Mourad, and Al-Ajlan, Monirah Abdullah. "Deep Learning Algorithm for Cyberbullying Detection." *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 9, No. 9, 2018.
- [6] K. S. Alam, S. Bhowmik, and P. R. K. Prosun. "Cyberbullying Detection: An Ensemble-Based Machine Learning Approach," *Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, pp.
- [7] Teoh, Hwai Teng, and Varathan, Kasturi Dewi. "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches." *IEEE Access*, vol. 11, 2023, pp. 55533–55560.
- [8] Waheed, T., Zafar, A., Wasim, M., Ali, A., Ali, H., Shah, Z., & Atif, A. (2024). Cyberbullying Detection and Abuser Profile Identification on Social Media for Roman Urdu. *IEEE Access*, 12, 123339-123351.
- [9] J. Eronen, M. Ptaszynski, F. Masui, A. Smywiski-Pohl, G. Leliwa and M. Wroczynski, "Improve classifier training efficiency for automatic cyberbullying detection with feature density", *Inf. Process. Manage.*, vol. 58, no. 5, pp. 102616, 2022.
- [10] R. Annamalai, S.J. Rayen and J. Arunajsmine, "Social media networks owing to disruptions for effective learning", *Procedia Computer Science*, vol. 172, pp. 145-151, 2020.
- [11] K.S. Alam, S. Bhowmik and P. R K Prosun, "Cyberbullying detection: An ensemble-based machine learning strategy", *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, pp. 710-715, 2021.
- [12] N. M. G. D. Purnamasari, M.A. Fauzi and L.S. Dewi, "Cyberbullying identification in Twitter using support vector machines and information gain-based feature selection", *Indones J. Electr. Eng. Comput. Sci.*, vol. 18, pp. 1494-1500, 2020.
- [13] A. Perera and P. Fernando, "Accurate Cyberbullying Detection and Prevention on social media", *Procedia Computer Science*, vol. 181, pp. 605-611, 2021.
- [14] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, et al., "A nature-inspired-based approach for automated cyberbullying classification on multimedia social networking", *Math. Problems Eng.*, 2021.
- [15] Ranjan S, Sanket S, Singh S, Tyagi S, Kaur M, Rakesh N, Nand P. OCR based Automated Number Plate Text Detection and Extraction. In 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom) 2022 Mar 23 (pp. 621-627). IEEE.