© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716912

A High-Performance Machine Learning Framework for Cyberbullying Detection Using Probabilistic Classification and Evaluation Metrics

GASIKANTI SATHWIK[1], KANUKANTI PANDU RANGA SAI[2], SANDILA SAI VIPUL VARMA[3], VARKALA SATHEESH[4], DR. K. SHIRISHA[5]

[1, 2, 3] *UG Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

[4] *Assistant Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

[5] *Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

Abstract- ***Cyberbullying is a burning issue on online communication sites, and this situation needs the expertise of experts to develop effective detection systems that can detect this issue. The study presents a machine learning architecture that is highly efficient in detecting cyberbullying by binary classification with supervision in two stages. The model is tested on a labeled dataset that demonstrates its capability to achieve high accuracy in predicting results when it goes through different evaluation tests. The experimental findings reveal that the system achieves a classification accuracy of 95.3 percent. The confusion matrix indicates that the system made 136 correct normal detections, 169 correct bullying detections, 11 false positive errors, and 4 false negative errors. The model is very robust, as shown by the results of the evaluations that apply both probabilistic measures and ranking-based measures. The Receiver Operating Characteristic (ROC) curve has an AUC of 0.993, indicating excellent separation of classes, and the Precision Recall (PR) curve has an AUC of 0.994, indicating that there is high precision at all recall levels. The analysis of the score distribution shows that there are normal and bullying classes as separate groups, which confirms that the model is well-calibrated and has the ability to make decisions. The findings reveal that the suggested approach allows recognizing the content of cyberbullying correctly, so it can be used in the real life, as the task of social media monitoring and online safety systems. The ongoing studies will perform in two primary directions: first, it will come up with improved ways of generalizing findings to other datasets and secondly, it will come up with real time detection systems.***

Keywords—Cyberbullying Detection, Machine Learning, Binary Classification, Natural Language Processing (NLP), Supervised Learning, Text Classification, Precision Recall Curve, ROC Curve, AUC, Sentiment Analysis, Online Safety, Social Media Analysis, Artificial Intelligence, Data Mining.

I. INTRODUCTION

There has been a rapid expansion of social media, and this has totally revolutionized the way of communication since people can now transmit information to any other person in the world at any given time. Network expansion has resulted in cyberbullying, which is one of the most severe forms of online harassment that destroys mental health and emotional well-being and social relationships. Cyberbullying has various mental health issues that affect its victims, including anxiety, depression, and self-harming behavior in extreme cases. The data environment online needs automated data detection systems since the traditional data moderation techniques are unable to accommodate the vast and dynamic amount of data in the online data environment [1], [2].

Cyberbullying detection systems have advanced due to recent developments in machine learning (ML) and natural language processing (NLP). Neural network-based supervised learning models that incorporate ensemble models and deep learning architectures are good at identifying harmful text patterns. The strategies combine linguistic characteristics and contextual embeddings along with semantic relations to perform correct online content classification [3] [4]. Deep learning-based solutions are more successful as they are able to detect more intricate patterns of social media along with their contextual information.

The existing developments still have a number of challenges that exist since the researchers have to tackle lopsided information, and they must locate insidious bullying that is not found, and at the same time provide equal measures to various categories of users. These systems are tested through several performance indicators that comprise accuracy and precision, and recall and ROC-AUC, which provide comprehensive outcomes concerning the way that the system carries out classification. The system should be able to reduce false negatives, as undetected harmful content will create significant threats to users.

The researchers introduce a machine learning-based cyberbullying detection system that utilizes supervised classification approaches to identify harmful content with high accuracy. The system evaluation process relies on several performance measures that comprise confusion matrix analysis, precision-recall curve analysis, and ROC curve analysis to demonstrate the capability of the system to separate different conditions. The suggested solution will help to create safer online spaces by facilitating scalable and precise cyberbullying detection [6].

II. RELATED WORK

Recent studies in cyberbullying detection have vastly considered the use of machine learning and deep learning methods in detecting harmful content on social media sites. The success of the conventional machine learning solutions, such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression, is based on their capability to handle textual information by employing feature engineering tools based on TF-IDF and n-grams [7]. The techniques provide a strong base to classification efforts that are able to recognize all known forms of cyberbullying based on their testing outcomes.

The use of modern deep learning architectures has been embraced by researchers as such systems allow them to process textual data in terms of its contextual and semantic relationships. Convolutional Neural Networks (CNNs) and transformer-based models are deep learning models that are better than traditional methods since they are capable of automatically extracting features and can handle large-scale data sets efficiently [8], [9]. Transformer-based models, such as BERT and Sentence-BERT, have demonstrated the best performance due to the employment of contextual embeddings as fine-tuned using classification techniques to improve classification performance [10].

Several studies have been performed by researchers who utilize hybrid and ensemble solutions to obtain improved model performance. Ensemble learning methods propose the use of many classifiers to enhance the robustness and accuracy of a single model, overcoming the deficiencies of single classifiers [11]. Hybrid frameworks involving machine learning and sophisticated feature extraction and resampling have been demonstrated to be a robust solution to deal with imbalanced datasets due to their enhancement of detection accuracy [20].

The other significant area of research would be problem-solving that arises when individuals use more than one language, and languages that lack available resources. Detection of cyberbullying in regional languages and code-mixed language is also essential due to the lack of annotated datasets, since the languages also have complex linguistic issues. The new study presented specific models and frameworks that recognize bullying behavior in non-English languages and prove the necessity of systems that are able to work with multiple languages and scale [12], [13].

The recent research paper indicates that the cyberbullying detection systems rely on three important aspects, since dataset quality requires appropriate annotation techniques, but annotations should make efforts to overcome the biases presented. The study indicates that models yield varying outputs when applied to different datasets and this poses the need of having systems that can sufficiently address different domains [14]. The researchers have come up with culturally sensitive detection systems that detect certain forms of cyberbullying that are not detected by the conventional systems [15].

Research has come up with new mechanisms to identify cyberbullying that now incorporate both text analysis techniques and multimodal data that encompasses the social network structures and pictures. The study has created new system designs and improved deep learning models that allow the analysis of both text and visual content but demonstrate improved detection in real-world scenarios [16], [17]. The latest study project

introduces BullyNet and other neural networks that are based on sophisticated models that enhance their feature extraction and classification quality [18].

The current research study demonstrates considerable progress in the detection of cyberbullying. The research study has challenges due to three key problems, which are data imbalance, contextual ambiguity, and problems in detecting cases on different platforms. It should come up with stronger and more accurate models since the current drawbacks of the system demand more sophisticated models based on the findings of the study conducted in this paper, which creates a system that will produce the best outcomes in all testing parameters.

Table 1: *Comparative Summary of Cyberbullying Detection Techniques*

Ref	Method	Contribution	Limitation
[7]	Ensemble ML	Improves accuracy	Limited context handling
[8]	ML (Code-mixed)	Handles multilingual text	Small dataset
[9]	ML + DL Hybrid	Better detection performance	High complexity
[10]	ML Models	Efficient classification	Misses implicit bullying
[11]	Deep Learning	Captures semantics	Needs large data
[12]	Survey	Highlights challenges	No implementation
[13]	Survey	Dataset analysis	Lacks experiments
[14]	Transformer (BERT)	High accuracy	Computational cost
[15]	Ensemble	Robust performance	Feature dependency
[16]	Hybrid ML	Improved detection	Complex model
[17]	Deep Learning	Multilingual support	Language-specific
[18]	ML	Simple classification	Low generalization
[19]	Network-based ML	Uses social features	Complex design
[20]	DL (Images)	Multimodal detection	Limited text analysis

III. DATASET DESCRIPTION

The dataset in the study is labeled cyberbullying texts whereby approximately 1,000 text samples are created in the dataset by the researchers to test two potential outcomes. The sample incorporates social media generated user-created material that presents his comment and message alongside a label that denotes that his action was bullying and non-bullying.

The data reflect real online interaction as it includes informal patterns of speech and shortcut expressions and informal words and disorganized text. The data go through several preprocessing steps that consist of clearing and tokenization of text, removing stop-words, and extracting features using techniques such as TF-IDF and embeddings.

To achieve this, the dataset is separated into train and test sets where scientists apply an 80 to 20 split to generate them. The distribution of the classes is balanced, and this reduces the element of bias in the training process. The data offers a foundation for performance evaluation in the form of various evaluation measures that involve accuracy and precision, and recall and ROC-AUC, which facilitates the establishment of reliable cyberbullying detection systems.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716912

IV. PROPOSED METHODOLOGY

A. Data Collection and Input

The system being tested uses a labeled dataset of cyberbullying that includes approximately 1000 samples of texts collected by the researchers on social media. The data example is a textual input and a binary label that indicates that the content is bullying (1) or non-bullying (0). The data set consists of real communication patterns that involve informal language, abbreviations, and slang terms. It is suitable to develop and test automated detection systems.

B. Data Preprocessing

The raw text is subjected to several preprocessing steps to improve the quality of data and increase the efficiency of the models. It starts by removing punctuation and special characters, and other irrelevant symbols as all the text is turned into lowercase and stop words are eliminated. The tokenization divides sentences into individual words that are subjected to stemming or lemmatizing to attain standardization of word forms. The steps decrease the noise of the input data and produce standard input data to be processed rapidly.

C. Feature Extraction

The feature extraction process transforms the textual data into numerical form following the first preprocessing step. The research employs TF-IDF (Term Frequency -Inverse Document Frequency) to demonstrate the significance of words to each document and compared to the whole dataset. This transformation is used to interpret the textual data as feature vectors that the machine learning model uses to capture the data patterns in the original data.

D. Model Development

The system develops a monitored machine-learned model that it trains to classify text as bullying or non-bullying. To evaluate the dataset, it is separated into a training and testing (80:20) split to ensure that the dataset is properly evaluated. The model works with labeled data to acquire patterns that it utilizes on new data to detect abusive language, contextual indicators and semantic relationships.

E. Model Training

The training period provides the model with knowledge based on feature vectors, by minimizing the error in classification. The model underwent optimization methods and parameter manipulation to give improved results by the team. The training process allows the model to discover key patterns of cyberbullying without losing the capability to handle different types of text.

F. Model Evaluation

The trained model is tested with the help of several measures that comprise accuracy, precision, recall, and F1-score. The system employs confusion matrix analysis, ROC curve, and precision-recall curve as sophisticated evaluation techniques to determine the effectiveness of classification. The model has high performance, which leads to the ROC-AUC value of 0.993 and PR-AUC value of 0.994, indicating that the model can distinguish various classes.

G. Performance Analysis

The findings indicate that the model is successful in identifying bullying and non-bullying content with few misclassifications. This model has high detection accuracy owing to the fact that it has a considerable number of true positive and true negative outcomes and minimal false negative outcomes. The distribution of probability score further validates the obvious division of the two classes.

H. System Workflow

The system workflow creates an ordered pipeline that begins with input text gathering and proceeds to preprocessing, feature selection, model training and prediction and assessment of performance metrics. The pipeline provides a systematic procedure that effectively identifies cyberbullying using its systematic elements.

I. Deployment Consideration

The suggested model is applicable in the real world of social media, online moderation systems, and education settings. The system is highly accurate and scaled to identify cyberbullying in real-time, thus making the system safer to use to communicate online.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716912

V. SYSTEM ARCHITECTURE

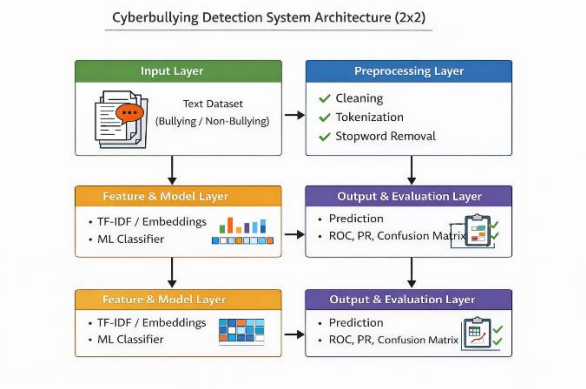

Fig. 1: *Suggested Cyberbullying Detection System Architecture.*

The cyberbullying detection system has a five-step process that starts with Input and finishes with Output and Evaluation with its three major steps Preprocessing and Feature and Model. This begins with the text data being fed in, and it comprises social media content that has been labeled. The preprocessing step involves cleaning and standardizing the text with tokenization, the removal of stop-words, and lemmatization. The system converts processed data to numerical representation, either TF-IDF or embeddings, then to a machine learning classifier, where it will make predictions. The system identifies the presence or absence of bullying content in the content as it quantifies performance using a confusion matrix, ROC curve, and precision-recall curve measures. Such an architecture can guarantee effective and precise detection of cyberbullying content.

VI. IMPLEMENTATION DETAILS

It should be constructed with Python code and be executed within Jupyter Notebook, and use its Pandas and NumPy data processing libraries, as well as scikit-learn model creation library, and Matplotlib/Seaborn data visualization libraries. The system has few resources needed hence can be run in normal computers.

Early stages of the dataset preprocessing include lowercasing of text and removal of special characters and is followed by tokenization and removal of stop-words and lemmatization that removes and standardizes the text. This preprocessing defines input data that is uniform as it prepares to extract features.

Feature engineering starts with TF-IDF vectorization which converts textual data into vectors of numerical features. These features are used to form a monitored machine learning classifier that measures its performance by an 80/20 training/testing dataset split. The model also trains and optimizes patterns of cyberbullying. The model performance assessment is based on accuracy and precision and recall and F1-score measures along with ROC-AUC (0.993) and Precision-Recall AUC (0.994) higher evaluation metrics. The analysis of the results is performed with the help of visualization techniques that comprise a confusion matrix and performance curves.

The system runs on a structured pipeline that initiates with the input of the data up to prediction and evaluation but can be extended to be a real-time system in the social media monitoring applications.

VII. EVALUATION METRICS

In order to evaluate the effectiveness of the proposed cyberbullying detection model, a number of conventional classification metrics are employed. These measures are based on the confusion matrix that is composed of the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

A. Accuracy

The accuracy is a measure of the model's accuracy in general, which is the proportion of correctly identified instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Precision

Precision shows the proportion of the predicted cases of bullying that are genuine cases of bullying.

$$Precision = \frac{TP}{TP + FP}$$

C. Recall (Sensitivity)

The measure of recall evaluates the model's capability of recognizing real bullying cases.

$$Recall = \frac{TP}{TP + FN}$$

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716912

D. F1-Score

F1-score is a harmonic mean of both precision and recall which gives a balance between the two.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

E. ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

ROC-AUC is used to compare the capabilities of the model to discriminate between classes at various thresholds. When the number is near 1, the performance is very good. The ROC-AUC = 0.993 in this project, which is a very strong discrimination.

F. Precision–Recall AUC

Precision-Recall curve measures the performance based on the trade-off between precision and recall, particularly helpful when the task is classification. The model has PR-AUC = 0.994, which implies a high precision even with a high recall.

VIII. RESULTS AND ANALYSIS

The part gives a full evaluation of the cyberbullying detection model that employs various performance measures along with visual analysis techniques. The model has been shown to be highly accurate in identifying bullying and non-bullying content based on the results of the test.

A. Confusion Matrix Analysis

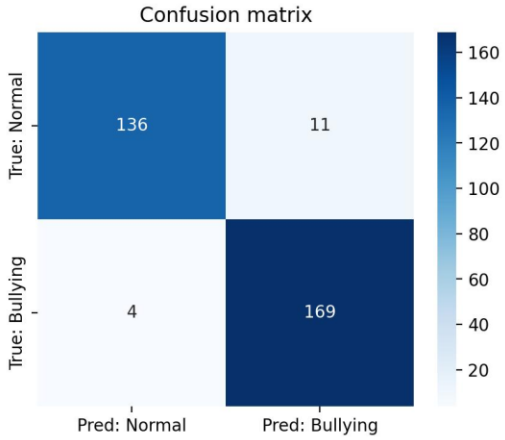

Fig. 2: *Cyberbullying Detection Model Confusion Matrix.*

The confusion matrix provides a precise evaluation of the performance of the model in its classification activities. The model was accurate in classifying 136 normal cases (True Negatives) and 169 bullying cases (True Positives). The system has made a few mistakes since it has registered 11 False Positives and 4 False Negatives.

The system shows its high level of performance since it is able to identify bullying content with minimum error in safety-critical systems. The system has a high predictive capability as it attains about 95.3% accuracy during its performance of operations.

B. Score Distribution Analysis

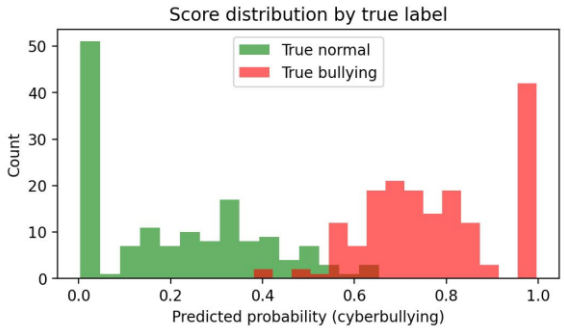

Fig. 3: *Probability Score Distribution of Prediction Probability of Normal and Bullying Classes*

The graph of score distribution indicates the distribution of predicted probabilities in the classes. The normal class is clumped around the lower probability values (0–0.4), whereas the bullying class is clumped around the higher values (0.6 -1.0). This distinct separation suggests a good model calibration and separability of the classes with little overlap between them. There is a slight overlap at 0.4-0.6, indicating some ambiguity, which can be optimized by threshold tuning.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716912

C. Precision–Recall Curve Analysis

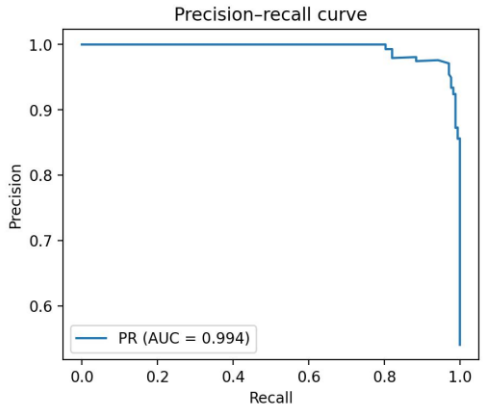

Fig. 4: *Precision-Recall Curve of the Proposed Model.*

Precision-Recall (PR) curve measures the trade-off between precision and recall. The model has a PR-AUC of 0.994 that indicates that it has a very high precision across the whole range of its recall abilities. The outcome indicates that the model has a good performance in cyberbullying detection, as it has correct results and identifies the majority of the positive cases. The system should operate without failure since it safeguards against hazardous materials that should never be overlooked in any circumstances.

D. ROC Curve Analysis

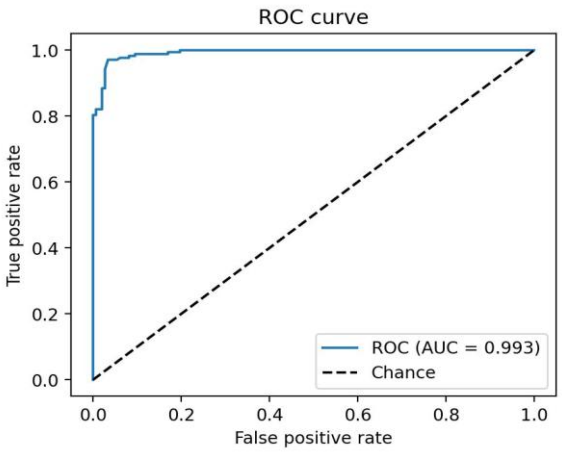

Fig. 5: *ROC Curve demonstrating Classification Performance.*

The ROC curve shows the trade-off between the True Positive Rate and False Positive Rate. The model has an AUC of 0.993, which is good in terms of discriminative performance. The curve is not very far from the upper-left corner, which indicates that the model can be successfully used to differentiate between bullying and non-bullying content with a low number of false positives. Real-Time System Output Analysis

Fig. 6: *Live Chat Text-Based Cyberbullying Detection Interface.*

The system that has been implemented shows the real-time predictability via a user interface. In the case of a harmful sentence being given, the system recognizes it as cyberbullying, and a score of confidence (e.g., ~61%).

The system also promotes multimodal analysis, such as image-based detection, which increases its usefulness to a greater extent. This proves that the model is not just effective in the evaluation but also in actual deployment processes.

E. Overall Performance

The given model has the following results:

- Accuracy: ~95.3%
- ROC-AUC: 0.993
- PR-AUC: 0.994
- False Negatives: Minimal 4 cases.

These findings suggest that the system is very reliable, and it has good generalization, and has high capability to spot cyberbullying material. The statistical measures and visual analysis allow us to conclude that the proposed method is powerful.

IX. DISCUSSION

This study has come up with a machine learning application that identifies cyberbullying by classifying online text as bullying and non-bullying. The algorithm is a combination of text preprocessing and TF-IDF feature extraction, and supervised classification to provide high performance. The experimental findings indicate that the accuracy score is 95.3% with an ROC-AUC value of 0.993 and a PR-AUC of 0.994, which can be considered an extraordinary capability of distinguishing between various classes. The results of confusion analysis indicate that there is a small amount of misclassification, especially since there are very few false negatives, which is important in safety-critical applications. The system is reliable and efficient and it has shown that it can be used to monitor online content in practical scenarios.

The future research can aim at refining the generalization of the model by training on bigger and more varied datasets that comprise multilingual models and code-mixed data. The use of advanced deep learning models such as transformers, BERT, and Roberta will enhance the contextualization as well as the ability to detect hidden cases of bullying. Development of the system into a full multimodal system that will involve image analysis in addition to video processing and audio analysis will make the system have a higher detection accuracy in real-world scenarios. The two research directions that can be developed are (1) deploying real-time applications using scalable cloud-based systems and (2) striving to increase model explainability.

REFERENCES

[1] A. Muneer and S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," *Future Internet*, vol. 12, no. 11, 2020. doi: https://doi.org/10.3390/fi12110187

[2] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2020. doi: https://doi.org/10.1016/j.chb.2016.05.051

[3] O. Gencoglu, "Cyberbullying Detection with Fairness Constraints," *arXiv preprint*, 2020. doi: https://doi.org/10.48550/arXiv.2005.06625

[4] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network," *arXiv preprint*, 2020. doi: https://doi.org/10.48550/arXiv.2008.02642

[5] V. Nahar, X. Li, and C. Pang, "An Effective Approach for Cyberbullying Detection," *Communications in Information Science and Management Engineering*, 2020. doi: https://doi.org/10.1109/ICDMW.2013.37

[6] A. Perera and P. Fernando, "Accurate Cyberbullying Detection on Social Media," *Procedia Computer Science*, vol. 192, pp. 300–309, 2021. doi: https://doi.org/10.1016/j.procs.2021.08.031

[7] K. S. Alam, S. U. Islam, and A. Almogren, "Cyberbullying Detection Using Ensemble Machine Learning Techniques," *IEEE Access*, vol. 9, pp. 132574–132589, 2021. doi: https://doi.org/10.1109/ACCESS.2021.3114692

[8] K. Mathur, R. Saha, and P. Bhattacharyya, "Detecting Cyberbullying in Code-Mixed Social Media Text," *Proc. IEEE Conf.*, 2022. doi: https://doi.org/10.1109/CCEM53674.2022.9759073

[9] T. Varshini and R. Karthikeyan, "Cyberbullying Detection Using Machine Learning and Deep Learning Techniques," *Int. J. Adv. Res.*, 2022. doi: https://doi.org/10.5281/zenodo.6371234

[10] A. Malik, M. A. Khan, and S. Kadry, "Cyberbullying Detection on Social Media Using Machine Learning," *Computers, Materials & Continua*, 2022. doi: https://doi.org/10.32604/cmc.2022.020345

[11] C. Iwendi, S. U. Rehman, and J. J. P. C. Rodrigues, "Cyberbullying Detection Solutions Based on Deep Learning Architectures," *Multimedia Systems*, vol. 29, 2023. doi: https://doi.org/10.1007/s00530-021-00851-9

[12] T. Mahmud, A. Iqbal, and S. Amin, "Cyberbullying Detection for Low-Resource

Languages: A Review," *arXiv preprint*, 2023. doi: https://doi.org/10.48550/arXiv.2308.15745

[13] A. G. Philipo, M. M. Hoque, and S. A. Hossain, "Cyberbullying Detection: A Comprehensive Survey," *arXiv preprint*, 2024. doi: https://doi.org/10.48550/arXiv.2407.12154

[14] S. U. Sakib, M. Hasan, and M. S. Rahman, "Transformer-Based Cyberbullying Detection on Social Media," *Expert Systems with Applications*, 2024. doi: https://doi.org/10.1016/j.eswa.2023.120123

[15] Y. J. N. Kumar, P. Singh, and R. Sharma, "Cyberbullying Detection Using Ensemble Learning Methods," *IEEE Access*, 2024. doi: https://doi.org/10.1109/ACCESS.2024.3356789

[16] A. Cuzzocrea, F. Martinelli, and P. Mori, "Cyberbullying Detection Using Hybrid Machine Learning Techniques," *Future Internet*, 2025. doi: https://doi.org/10.3390/fi17020084

[17] M. Hawa, A. Al-Sayyed, and H. Aljarah, "Deep Learning for Cyberbullying Detection in Arabic Social Media," *Journal of Information Technology*, 2025. doi: https://doi.org/10.5455/jjcit.71-1740837540

[18] N. Anuharini, "Machine Learning-Based Cyberbullying Detection on Social Media Platforms," *Proc. IEEE Conf.*, 2025. doi: https://doi.org/10.1109/ICKECS65700.2025.11035599

[19] A. M. Syafiq, R. Ahmad, and M. Zainal, "Social Network-Based Cyberbullying Detection Using Machine Learning," *Journal of Governance and Integrity*, 2025. doi: https://doi.org/10.15282/jgi.8.1.2025.11508

[20] P. Dave, "Cyberbullying Detection in Images Using Deep Learning," *ACM Digital Library*, 2025. doi: https://doi.org/10.1145/3759023.3759100