IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyberbullying Detection Using Ml

An Intelligent Approach to Identifying Harmful Online Behavior

[1]Shrushti Dilip Adhave, [2]Sakshi Satish Shinde, [3]Vinay Gajula, [4]Suchit Nangre

[1,2,3,4]Student, Department of Computer Engineering ,Terna Engineering College.

[5]Dr.Ramesh Shahabade,

[5]Professor, Department of Computer Engineering,Terna Engineering College.

Abstract: With the rise of social media, online communication has become more accessible, but it has also led to a surge in cyberbullying where individuals face abusive and harmful content. The anonymity and reach of digital platforms have made it easier for such behavior to spread, posing serious mental health risks.This project presents a real-time cyberbullying detection system using Natural Language Processing (NLP) and Machine Learning (ML). Focusing on English, Marathi, and Hinglish, we collected live data via APIs and web scraping, followed by preprocessing and TF-IDF-based feature extraction. Models such as Random Forest, Logistic Regression, XGBoost, and Linear SVC were evaluated, with the Linear SVC and a stacked model performing best.The system detects and filters offensive content in real-time and features a user-friendly interface for both users and administrators, supporting safer and healthier digital interactions.

***Index Terms* - Cyberbullying, ML Model, Accuracy**

I. Introduction

In today's digitally connected world, the rise of social media and online communication platforms has transformed the way people interact. While these platforms offer numerous advantages, they have also introduced new challenges—most notably, cyberbullying. Cyberbullying involves using digital means to harass, threaten, or demean others, often under the shield of anonymity. This behavior can have severe psychological impacts, especially on teenagers and young adults. As the prevalence of cyberbullying increases, the need for intelligent, real-time detection systems becomes more urgent.

Various solutions in the form of third-party applications have been deployed but the problem with these applications is that they are based on a simple keyword matching technique which gives less accurate results. Manually adding data in the database could be a subjective point of view of the creator and majority of the drawbacks include lack of labeled database or using a biased data set. Few implementations include lexicons, which is a common way to use databases for the detection of abusive words. However, it limits the scope of the application and disregards the statements which may not use abusive words but have hurtful meaning[2].

To address these limitations, this project proposes a real-time cyberbullying detection system leveraging Machine Learning (ML) and Natural Language Processing (NLP). By analyzing live social media data across English, Marathi, and Hinglish languages, the system aims to identify harmful content dynamically. Unlike traditional methods, our approach uses trained models on diverse datasets to detect not only overtly abusive words but also contextually harmful language. Additionally, the system includes alert mechanisms to notify users, guardians, or administrators, helping to reduce the impact of cyberbullying before it escalates.

The paper is organized as follows. Section II describes Related work, Section III describes methodolog, The results and discussion are mentioned in Section IV, Section V describes challenges, Section VI includes comparison for different Models, Section VII is conclusion.

II. RELATED WORK

In [1] the authors focus on a machine learning-based methodology for detecting phishing websites, which pose a significant threat by tricking individuals into providing sensitive information. The study employs several machine learning algorithms, including Extreme Gradient Boosting, Decision Trees (DT), Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) to differentiate between legitimate and phishing URLs. The researchers used two well-known datasets, PhishTank and UCI, and applied K-fold cross-validation, feature selection, and hyperparameter tuning to enhance model performance. The Random Forest algorithm outperformed others, achieving 98.80% accuracy on the PhishTank dataset and 97.87% on UCI. Additionally, the model achieved 99% precision, recall, and F1-score, with a high AUC-ROC value of 99.89%.

In [2], the authors explore innovative approaches to detect cyberbullying incidents across social media platforms such as Twitter. The study focuses on real-time data collection from Twitter, consisting of headlines, comments, and trending posts, followed by a labeling process to identify instances of cyberbullying and cyber aggression. The collected data is analyzed to uncover correlations between various features and the occurrence of cyberbullying. The aim of the project is to identify cyberbullying at its origin— while it is being drafted—by leveraging Machine Learning and Natural Language Processing (NLP) techniques. Several machine learning algorithms were employed for the classification and prediction of cyberbullying messages, especially in Hinglish (a hybrid of Hindi and English). Among these, the study highlights the effectiveness of Linear Support Vector Classifier (SVC) and Stochastic Gradient Descent (SGD), which demonstrated faster training and prediction times compared to other models. Logistic Regression and Random Forest classifiers were also tested, with Random Forest delivering the best accuracy and F1-score at 97.1% and 97.2%, respectively. While Random Forest achieved the highest overall performance, Linear SVC provided the best recall score (97.13%) and required significantly less time for training and prediction, making it a more efficient model in terms of computational speed. Despite Random Forest's superior accuracy, its longer training and prediction times present a trade-off when compared to the faster, yet still highly accurate, Linear SVC model.

In [3] the authors focus on the issue of cyberbullying on Twitter. supervised machine learning techniques, particularly Support Vector Machines (SVM), applied to data which was collected by using Twitter Advanced Search. The study includes preprocessing steps such as emoticon conversion and URL removal to improve data quality, and dimensionality reduction techniques like PCA and LSA to streamline analysis. The SVM algorithm was successfully implemented for classifying cyberbullying content on social media platforms like Twitter. result indicates that the SVM based method achieves best accuracy and the performance improves if user specific data can be included. Due to high dimensional input space, few irrelevant features and linearly separable nature of text dataset, SVM performs better than other classification classification. algorithms for text.

In [4],The paper focuses on the increasing prevalence of cyberbullying due to the widespread use of social media. With many people using these platforms to troll and defame others, detecting and preventing such harmful comments has become essential. The proposed solution uses an ensemble learning approach to classify and detect cyberbullying comments. The model utilizes a voting classifier, combining Support Vector Machine, Logistic Regression, and Perceptron models, with the final prediction based on the majority vote. This approach achieves a 94% accuracy rate in detecting cyberbullying. The study collected over 30,000 datasets from platforms such as Kaggle, Twitter, and YouTube. After text preprocessing and feature extraction, the individual classifiers— Perceptron, Support Vector Machine, and Logistic Regression—were evaluated, followed by the creation of the voting classifier. The ensemble method outperformed the individual models, with a 94% accuracy. This approach enhances real-time detection of cyberbullying and promotes safer use of social media.

In [5],The paper focuses on detecting cyberbullying in social media using natural language processing and supervised machine learning. It identifies key cyberbullying themes like racism, sexual content, and abusive language. The proposed system adapts to evolving language and excludes sarcastic text from detection. Achieving 74.50% accuracy, with 74% precision, recall, and F1-score, the research aims to improve accuracy further. The paper emphasizes the serious impacts of cyberbullying and highlights the need for prevention. It critiques existing solutions for not fully considering language evolution and presents a system that uses Support Vector Machines and Logistic Regression, along with feature extraction techniques like TF-IDF and sentiment analysis, for more accurate detection.

In [6], the author highlights the exponential increase in social media users and how cyberbullying has emerged as a form of bullying through electronic messages. Social networks provide a rich environment for bullies, making users vulnerable to attacks. The author explores suitable actions for the prevention and detection of cyberbullying patterns using machine learning algorithms. A supervised learning approach was employed in the study. The results show that the Neural Network model performs better, achieving an accuracy of 92.8%, while the Support Vector Machine (SVM) achieves 90.3%. The model was evaluated using two classifiers: SVM and Neural Network.

In [7],The survey paper focuses on the pervasive issue of cyberbullying in online social media, particularly emphasizing two key characteristics: repetitive behavior and power imbalance. The authors define a Social Media Session-based Cyberbullying Detection (SSCD) framework that comprises four components, specifically designed to tackle the unique challenges posed by session-based cyberbullying detection, which differs from single text-based approaches.The paper reviews existing research on cyberbullying detection, with a focus on session-based detection. It explores both data and methodological perspectives and highlights the current limitations in model and dataset creation. Notably, while the defining characteristics of cyberbullying—repetition and power imbalance—are often considered during the annotation stage, their integration into model design remains in its early stages. The authors also conduct benchmark experiments comparing the performance of state-of-the-art models and large pre-trained language models across two datasets. Based on their findings, they propose best practices for creating session-based cyberbullying datasets and identify open challenges for future research, encouraging further development in fine-grained detection methods to better understand the nature of cyberbullying.

III.METHODOLOGY

Proposed methodology for cyberbullying detection using machine learning is shown in Fig 1. The process starts with data collection from social media platforms. The data undergoes dataset processing, which includes handling missing values, normalization, and oversampling to balance the classes. This is followed by data cleaning and splitting the dataset into training and testing sets.

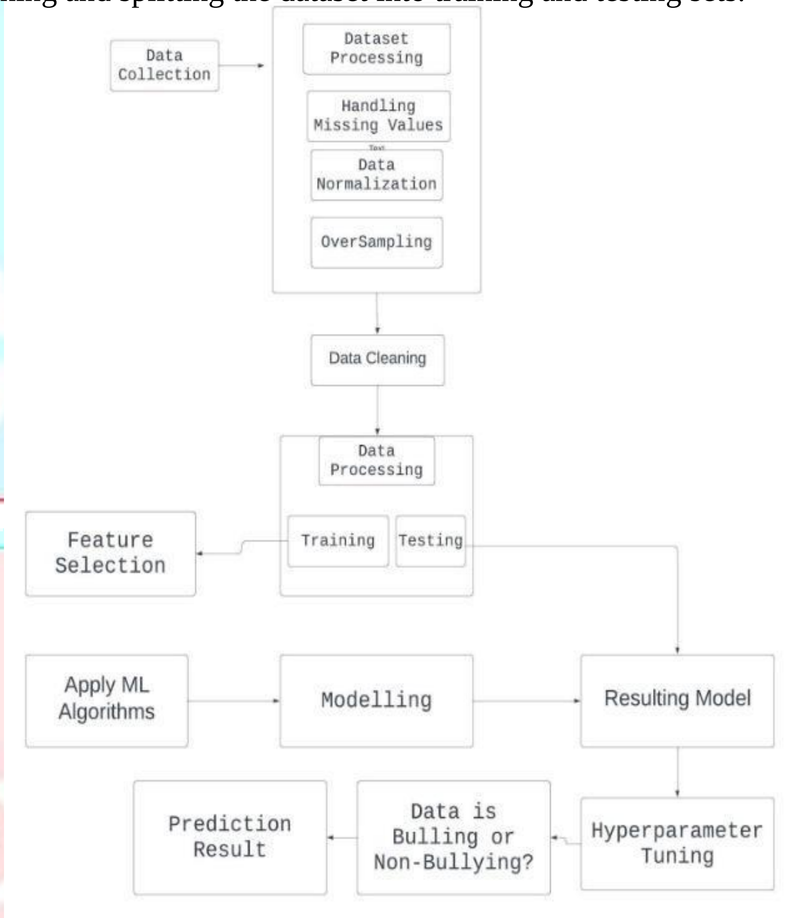

Fig 1: Proposed Methodology

Next, feature selection techniques are applied to extract important features, which are then used for model building. Various machine learning algorithms are applied to classify the data as bullying or non-bullying. The resulting model is 3 evaluated using the testing data, followed by hyperparameter tuning to improve accuracy. The final model is integrated into a real-time chat interface where bullying messages are blocked, and a warning is displayed, while non-bullying messages are allowed to appear in the chat.

3.1 Dataset Description

We utilized a dataset collected by scraping comments from various social media platforms. The dataset contains a total of 18149 instances and includes two key columns: Comments and Label. The comments are in three languages: English, Hindi, and Hinglish. The Label column contains binary values, where 0 represents non-bullying comments, and -1 represents bullying comments. The dataset was divided into two parts: 80% for training and 20% for testing. The multilingual nature of the data ensures robustness in detecting cyberbullying in a diverse linguistic context.

3.2 Proprocessing the Data

The scraped data was unstructured — filled with random symbols, links, and noise. Therefore, we implemented the following preprocessing techniques:

Cleaning the Text:
We removed unnecessary elements such as URLs, emojis, and special characters to provide clean, structured input for the model.

Tokenization and Lemmatization:
We broke the text into individual words (tokens) and applied lemmatization to convert them into their root forms (e.g., "running" became "run"). This step simplified the language while preserving its meaning.

Stop Word Removal:
Common words like "the," "is," and "and" were filtered out, allowing the model to concentrate on more meaningful cyberbullying.

Handling Imbalanced Data:
Since cyberbullying posts were underrepresented, the dataset was imbalanced. To mitigate this, we applied techniques such as SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic data for the minority class, along with under-sampling the majority class to balance the dataset.

Feature Extraction:
To make the cleaned text data interpretable by machine learning models, we converted the text into numerical format:

TF-IDF (Term Frequency-Inverse Document Frequency):
We employed TF-IDF to weigh words based on their uniqueness and importance in identifying bullying-related content, reducing the influence of overly common terms.

Word Embeddings:
We also utilized pre-trained word embeddings such as Word2Vec to understand semantic relationships between words. These helped the model capture the context and deeper meanings often hidden in slang or subtle cues in text.

Accuracy measures how often the model correctly predicts the outcome, calculated as the ratio of correctly classified instances (true positives and true negatives) to the total instances.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision measures the accuracy of positive predictions. It tells us what proportion of the instances predicted as positive (e.g., bullying) are actually positive. High precision means fewer false positives, which is important when the cost of a wrong positive prediction is high.

Precision = TP / (TP + FP)

Recall (also called sensitivity) measures the model's ability to find all actual positive cases. It tells us what proportion of actual positives were correctly identified by the model. High recall is 5 important when missing a positive case (false negative) has serious consequences.

Recall = TP / (TP + FN)

F1 Score is the harmonic mean of precision and recall. It gives a single score that balances both concerns, especially useful when there is an uneven class distribution or when both false positives and false negatives are important to minimize.

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

ROC (Receiver Operating Characteristic) curve is a graphical representation of a model's performance, plotting the true positive rate (sensitivity) against the false positive rate. The AUC (Area Under the Curve).

IV. RESULTS AND DISCUSSION

This chapter presents the performance evaluation.

TABLE 2: COMPARISON FOR MODEL

Algorithms	Accuracy	Precision	Recall	F1 score	Prediction time
XGBoost	0.78	0.81	0.78	0.77	60.79
SGD Classifier	0.86	0.86	0.86	0.86	0.23
Random Forest	0.84	0.84	0.84	0.84	136.46
Stack Model	0.87	0.87	0.87	0.87	0.02

As shown in Table 2, Random Forest achieved the highest accuracy (0.9760) and F1 score (0.9813) but had a higher prediction time (1.80 sec). Linear SVC provided competitive accuracy (0.9723) with the fastest prediction time (0.002 sec), making it ideal for real-time use. Logistic Regression followed closely, while XGBoost and K Neighbours performed comparatively lower, with K Neighbours having the slowest prediction (12.89 sec).Considering both performance and speed, Linear SVC was chosen for deployment in the system for English and Hinglish language detection

TABLE 3: COMPARISON FOR MODEL

Algorithms	Accuracy	Precision	Recall	F1 score	Prediction time
XGBoost	0.78	0.81	0.78	0.77	60.79
SGD Classifier	0.86	0.86	0.86	0.86	0.23
Random Forest	0.84	0.84	0.84	0.84	136.46
Stack Model	0.87	0.87	0.87	0.87	0.02

Table 3 illustrates comparative analysis of XGBoost, SGD Classifier, Random Forest, and a Stack Model was conducted for cyberbullying detection in Marathi. The Stack Model outperformed all individual models, achieving the highest metrics (accuracy, precision, recall, F1 score = 0.87) with the lowest prediction time (0.02 sec).

While SGD Classifier showed competitive performance (0.86) and low latency (0.23 sec), Random Forest had high accuracy (0.84) but was impractical due to a high prediction time (136.46 sec).

XGBoost recorded the lowest performance (accuracy = 0.78). Considering both effectiveness and speed, the Stack Model was selected for real-time deployment in the Marathi cyberbullying detection module.

Chat Application Interface:

The below images showcase the real-time chat application developed for cyberbullying detection. The system supports multilingual input and integrates the deployed machine learning models to monitor conversations. It uses the Linear SVC model for detecting inappropriate content in English and Hinglish, and the Stacked Model for Marathi.

Fig.2 Welcome Page

Fig.3 Chat Window

As seen in the below image, whenever a user sends a message containing offensive or abusive language, the system immediately identifies the inappropriate content and displays a warning message: "Warning: Inappropriate content detected!" This ensures that users are alerted instantly, maintaining a safe and respectful communication environment.

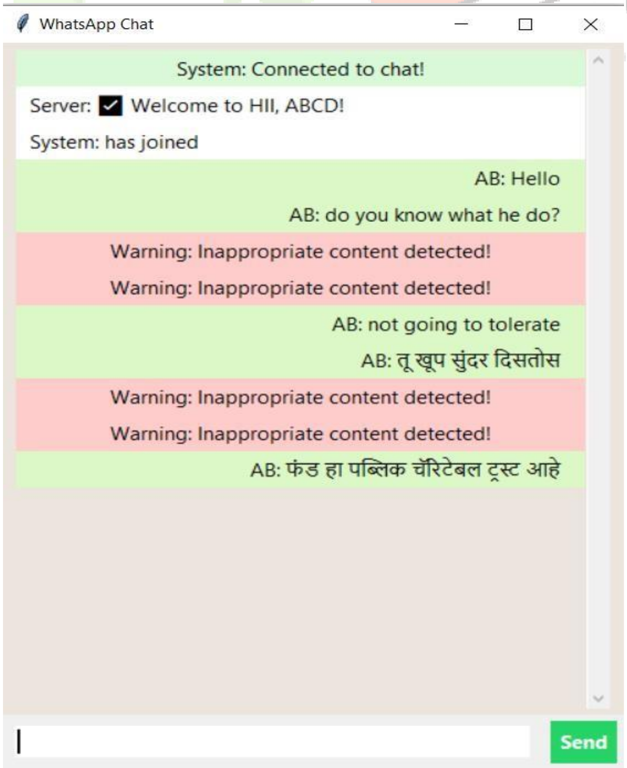

Fig.4 Working of chat window

The backend terminal (bottom section of the screenshot) confirms successful loading of all models and vectorizers, and logs each message exchanged, along with detection results. This illustrates the smooth interaction between frontend and backend modules, validating the system's ability to perform accurate and fast cyberbullying detection in a real-time chat setting.

Fig. 5 Backend Working

V. CHALLENGES IN CYBERBULLYING DETECTION:

Data Imbalance: The number of bullying-related comments is often much smaller than normal conversations, leading to a class imbalance problem.

Evolving Language: New slang, abbreviations, and emojis are constantly being created, making it difficult for static models to keep up with evolving cyberbullying language.

Contextual Understanding: Understanding context is essential as certain phrases may only be offensive in specific contexts (e.g., sarcasm).

Ethical and Privacy Concerns : The use of personal data from social media platforms raises privacy concerns, making it essential to anonymize and responsibly use datasets.

Resource Constraints in Real-Time: Deploying detection systems for real-time monitoring requires significant computational and storage resources, especially for high-traffic platforms.

Detection in Encrypted Spaces: With the rise of encrypted messaging platforms, detecting cyberbullying in private conversations while respecting privacy is a major challenge.

Temporal Relevance of Content: Some bullying-related comments lose their offensive nature over time due to changing social trends, making static datasets less effective for training models.

Cross-Platform Variability: Different social media platforms have unique structures, vocabularies, and interaction patterns, making it difficult to develop a unified detection model.

Classifying Indirect Bullying: Cyberbullying is not always direct; indirect forms like spreading rumors, excluding individuals, or posting harmful memes are harder to detect.

VI. COMPARISON FOR DIFFERENT MODEL

Paper Name	Models	Datasets	valuation Metrics	Advantages	Drawbacks
A Machine Learning Approach fo r Phishing Attack Detection.	Extreme Gradient Boosting, Decision Tre e, Logistic Regression, Random Forest (RF), and Support Vector Machine .	PhishTank and UCI dataset.	Accuracy : 98.80%, Precision:99 %, Recall: 99%, F1-Score:99% K-Fold CV::96.1% with the XGBoost model.	High Accuracy, Robust Performance Metrics, Comprehensive Feature Analysis, Utilization o f Multiple Datasets.	Overfitting Risks, Dependence on Quality of Data, Evolving Nature of Phishi ng Attacks.

Cyber-Bullying Detection in Hinglish Languages Using Machine Learning	Linear Support Vector Classifier, Decision Tree ,Naive Bayes, Random Forest, K-Nearest Neighbors (KNN), AdaBoost Classifier	Twitter Dataset, Hinglish extracting tweets from Twitter, chats from WhatsApp, and comments from YouTube, Crowdflower Datasets.	Random Forest best accuracy is 97.1% and F1 score-97.2% Linear SVC recall score-97.13%.	High Accuracy, Real-Time Detection, Diverse Feature Extraction.	Dependence on Labeled Data, Complexity of Language, Overfitting Risks, Limitations of Keyword Matching.
Cyber-Bullying Detection using Machine Learning Algorithms	Support Vector Machine (SVM), Naïve Bayes, Linear SVM, Tree-based Classifiers (J48).	Kaggle Dataset, Twitter Dataset, FormSpring.me Dataset.	AUC (Area Under the Curve): AUC score = 82%	Automated Detection, Feature Extraction, Human Validation	Insufficient Training Data, High Dimensionality, Dynamic Nature of Language.
Detecting the Presence of Cyberbullying using Machine Learning	Support Vector Machine (SVM), Neural Networks, Voting Classifier, Logistic Regression.	Kaggle.com,Twitter Dataset.	Accuracy=94%, For neural network = 91.76%, Support Vector Machine (SVM)= 89.87%.	High Accuracy, Ensemble Learning Benefits, Feature Extraction Techniques, Real-Time Detection.	Dependence on Data Quality, Complexity of Language, Overfitting Risks, Limited Contextual Understanding.
Accurate Cyberbullying Detection and Prevention on Social Media	Support Vector Machines (SVM), Logistic Regression.	Twitter Dataset.	Accuracy = 74.50%, Precision = 74%, Recall = 74%.	Comprehensive Feature Utilization, Comprehensive Feature Utilization, Integration of Multiple Detection Systems.	Limited Dataset for Training, Dependence on Language Evolution, False Positives and Negatives.

Social Media Cyberbullying Detection using Machine Learning	Support Vector Machine (SVM), Neural Network (NN).	Kaggle.com,Formspring.com datasets.	Accuracy: Neural Network (NN)=92.8%, Support Vector Machine (SVM)= 90.3%. F1-score: Neural Network (NN)=91.9%, Support Vector Machine (SVM)= 89.8%.	High Accuracy, N-Gram Models, Utilization of Advanced Techniques.	Dependence on Dataset Size, Imbalance in Data, Complexity of Language, Performance Variability.
Online Social Networks and Media	Rule-Based Classifier, CONcISE Algorithm, Time-Informed Gaussian Mixture Model (UCD), Latent Dirichlet Allocation (LDA), Graph Neural Networks, Deep Learning Models.	Social Media Extracted Dataset.	Accuracy: Higher Accuracy, Precision: Higher Precision.	Enhanced Understanding of Cyberbullying, Framework for Session-Based Detection, Diverse Feature Utilization, Addressing Class Imbalance.	Limited Dataset Quality, Insufficient Reporting Standards, Generalizability Limitations.

VII. CONCLUSION:

With the rapid expansion of social media, cyberbullying has emerged as a serious digital threat, affecting users across age groups, cultures, and regions. The primary motivation behind this project was to address this issue by building a **real-time cyberbullying detection system** capable of identifying and responding to offensive content online. What sets our system apart is its capability to understand and analyze content in multiple languages, including **English, Marathi**, and **code-mixed variations like Hinglish**. By integrating **machine learning** and **natural language processing** techniques, the system is trained to recognize informal, evolving, and multilingual online communication—making it highly relevant for the diverse Indian digital landscape. To showcase its real-time effectiveness, the system was implemented in a **custom-built WhatsApp-style chat application**, where offensive content is flagged or blocked instantly. From **data collection and preprocessing** to **model training and evaluation**, we successfully developed a robust, multilingual, and responsive prototype that fulfills our project goals.

VIII. References:

[1] T. Choudhary, S. Mhapankar, R. Bhddha, A. Kharuk, and R. Patil, "A Machine Learning Approach for Phishing Attack Detection," Journal of Artificial Intelligence and Technology, vol. 3, no. 3, pp. 108–113, 2023. doi: 10.37965/jait.2023.0197.

[2] K. Shah, C. Phadhtare, and K. Rajpara, "Cyber-Bullying Detection in Hinglish Languages Using Machine Learning," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 5, pp. 785-789, May 2022.

[3] M. Kini, A. Keni, Deepa, D. K. V., and D. C. H., "Cyber-Bullying Detection using Machine Learning Algorithms," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 5, pp. 690-695, May 2022.

[4] Arathi Unni ,Ranimol K R ,Linda Sebastian ,Rajalakshmi S ,"Detecting the Presence of Cyberbullying using Machine Learning ,"International Journal of Engineering Research & Technology (IJERT), vol. 3, May 2021.

[5] P. Fernando, "Accurate Cyberbullying Detection and Prevention on Social Media," International Journal of Computer Applications, vol. 1, no. 1, 2021.

[6] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, and E. Amer, "Social Media Cyberbullying Detection using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, 2019.

[7] P. Yi and A. Zubiaga, "Online Social Networks and Media," Information Processing, 2023.

[8] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-6.

[9] Azeez, Nureni & Idiakose, Sunday & Onyema, Chinazo & van der Vyver, Charles. (2021).Cyberbullying Detection in Social Networks: Artificial Intelligence Approach. Journal of Cyber Security and Mobility. 10. 745-774. 10.13052/jcsm2245-1439.1046.

[10] S. A. Oluwadare, O. A. Ayeni, and L. O. Ojuluwo, "Cyberbullying Detection Using K-Nearest Neighbour and Logistic Regression," Benin Journal of Physical Sciences, vol. 1, no. 2, pp. 117-130, Dec. 2024. [

11]S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 734-739.

[12]A. Alabdulwahab, M. A. Haq, and M. Alshehri, "Cyberbullying Detection using Machine Learning and Deep Learning," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10, pp. 424-432, Oct. 2023.

[13] V. Jain, V. Kumar, V. Pal and D. K. Vishwakarma, "Detection of Cyberbullying on Social Media Using Machine Learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1091- 1096.

[14] R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine International Learning," Conference on 2020 4th Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 297-301.

[15] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," in IEEE Access, vol. 9, pp. 88364-88376, 2021.

[16] A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber Bullying Detection on Social Media using Machine Learning," in ITM Web of Conferences, vol. 40, p. 03038, 2021