AI-Based Cyberbullying Detection and Prevention System Using Machine Learning, OCR, and Explainable AI

Priyanka Kasera[1], Saamin Sosan[2], Dr Renukadevi S[3]

[12]Student, School of Computer Science and Information Technology, Jain (Deemed-to-be University), Jayanagar, Bengaluru, Karnataka, India

[3]Assistant Professor, School of Computer Science and Information Technology, Jain (Deemed-to-be University), Jayanagar, Bengaluru, Karnataka, India.

Abstract

Cyberbullying has emerged as a critical issue in the digital era, significantly impacting individuals' mental health and well-being. Traditional detection methods rely on manual moderation and keyword filtering, which are often inefficient and lack contextual understanding. This paper presents an AI-based cyberbullying detection and prevention system that integrates machine learning, natural language processing, optical character recognition (OCR), and explainable AI techniques. A synthetic dataset of 10,000 entries was generated, containing labeled text data with attributes such as binary classification, category, severity, and emotional indicators. The system utilizes TF-IDF vectorization along with Logistic Regression, Random Forest, and One-vs-Rest classifiers to perform multi-level classification tasks. Additionally, OCR is used to extract text from images such as memes and screenshots. The system is deployed as a Streamlit-based web application that enables real-time detection, anonymous reporting, visualization dashboards, and mental well-being support. Experimental results demonstrate effective performance, highlighting the system's potential for real-world deployment.

Keywords: *Cyberbullying, NLP, Machine Learning, TF-IDF, OCR, Emotion Detection, Explainable AI, Streamlit, Mental Health*

1. Introduction

With the rapid growth of social media platforms and digital communication, cyberbullying has become a widespread concern affecting individuals across all age groups. Unlike traditional bullying, cyberbullying occurs in virtual environments, allowing perpetrators to remain anonymous while targeting victims repeatedly. This leads to severe psychological consequences, including anxiety, depression, and social withdrawal.

Existing solutions primarily focus on awareness and policy-based interventions, which are reactive rather than proactive. There is a growing need for intelligent systems capable of detecting harmful content in real time and providing preventive support mechanisms. In this context, artificial intelligence and machine learning offer promising solutions.

This research proposes a comprehensive system that not only detects cyberbullying but also incorporates features such as OCR-based image analysis, explainable AI, anonymous reporting, and mental well-being support, making it a holistic solution for digital safety.

2. Literature Survey

Several studies have explored cyberbullying from psychological, social, and technological perspectives.

Kowalski and Limber (2015) examined the psychological, physical, and academic effects of cyberbullying compared to traditional bullying. Their study revealed that cyberbullying leads to higher psychological distress, though the research was limited to U.S. high school students. [1]

Hinduja and Patchin (2017) focused on prevention and policy strategies. Their findings emphasized that structured prevention programs significantly reduce cyberbullying risks. However, their work primarily addressed policy-level solutions rather than automated detection systems. [2]

Tokunaga (2019) conducted a comprehensive review of cyberbullying victimization studies. The research highlighted long-term mental health issues such as anxiety and depression but lacked implementation of technological solutions for detection. [3]

Smith (2022) analyzed behavioral patterns and statistical trends, showing a rise in cyberbullying with increased social media usage. The study stressed awareness but did not propose real-time detection mechanisms. [4]

Ali et al. (2024) introduced machine learning approaches for detecting cyberbullying content. Their models achieved over 85% accuracy; however, the study was limited to English datasets and lacked generalization across diverse languages. [5]

3. Research Gap

From the above studies, the following gaps are identified:

- Lack of real-time detection systems
- Limited integration of image-based analysis
- Absence of explainable AI

- No mental wellbeing support mechanisms
- Limited deployment in practical applications

This project addresses these gaps by developing a multi-functional AI system with real-time detection, OCR integration, explainability, and mental health support.

4. Proposed System

The proposed system is a comprehensive AI-based cyberbullying detection platform designed to analyze both textual and image-based content. It integrates multiple machine learning models to perform binary classification, category prediction, severity estimation, and emotion detection. In addition to detection, the system provides anonymous reporting functionality, allowing users to safely report harmful content without revealing their identity.

A key feature of the system is the inclusion of Optical Character Recognition (OCR), which enables the extraction and analysis of text from images such as memes and screenshots. This extends the detection capability beyond plain text and addresses a major limitation of existing systems. Furthermore, the system incorporates an explainable AI module that highlights the most influential words contributing to the prediction, improving transparency and user trust.

Another important component is the mental wellbeing module, which allows users to log their emotional state and receive supportive suggestions. This feature distinguishes the proposed system from traditional detection tools by addressing not only the identification of cyberbullying but also its psychological impact.

The following components are:

- Text-based detection
- Image-based detection (OCR)
- Category classification
- Severity prediction
- Emotion detection
- Anonymous reporting system
- Mental wellbeing module
- Admin analytics dashboard

5. System Architecture

The architecture of the proposed system is designed to handle multiple stages of cyberbullying detection in an efficient and scalable manner. The system begins with the input layer, where users can provide either textual content or upload images containing embedded text. In the case of image input, the Optical Character Recognition (OCR) module extracts textual information using Tesseract, enabling further processing. The extracted or input text is then passed through a preprocessing stage, where it is cleaned by removing noise such as URLs, special characters, and stopwords.

Following preprocessing, the text is transformed into numerical features using the TF-IDF vectorization technique, which captures the importance of words and phrases. These features are then fed into multiple machine learning models, each responsible for a specific task, including binary classification, category prediction, severity estimation, and emotion detection. The output layer presents the results to the user in an interpretable format, along with explainability insights highlighting key contributing words. Finally, the system stores reports in a SQLite database and visualizes trends through an admin dashboard.

The system architecture consists of:

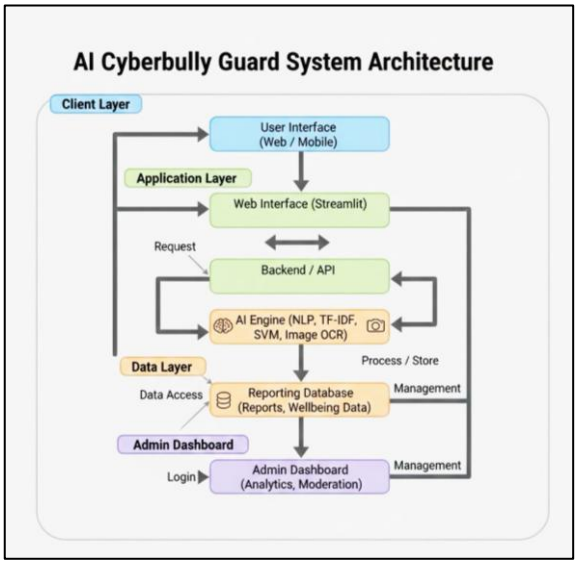

Fig.1 System Architecture Diagram

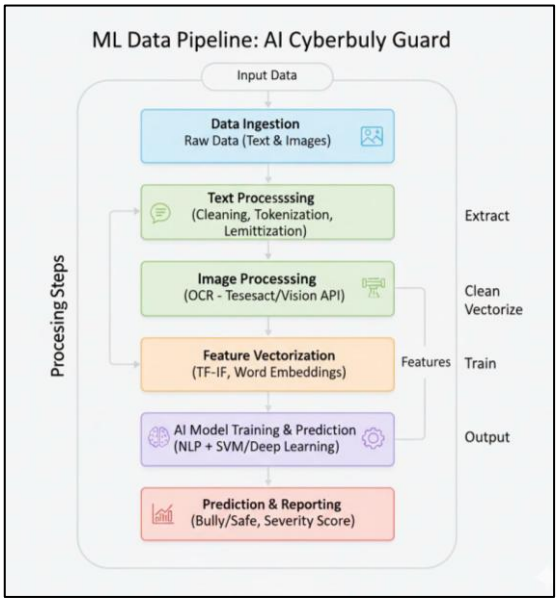

Fig.2 ML Data Pipeline Diagram

6. Methodology

The proposed system follows a structured pipeline beginning with dataset generation, preprocessing, feature extraction, and model training.

A synthetic dataset of 10,000 samples was created using predefined bullying and non-bullying phrases. Each entry includes attributes such as binary classification, category, severity score, emotional labels, and user identification.

Text preprocessing involves converting text to lowercase, removing noise such as URLs and special characters, and filtering stopwords. The cleaned text is then transformed into numerical features using TF-IDF vectorization, which captures the importance of words within the dataset.

Multiple machine learning models are trained for different tasks. Logistic Regression is used for binary classification due to its efficiency and interpretability.

Random Forest is used for category classification and severity prediction, while a One-vs-Rest approach is applied for multi-label emotion detection. Additionally, OCR using Tesseract is implemented to extract text from images, enabling the detection of cyberbullying in memes and screenshots.

6.1 Dataset Generation

The dataset used in this study consists of 10,000 synthetically generated samples, including both bullying and non-bullying text. Each entry contains attributes such as text content, binary classification labels, category, severity score, emotion tags, and user identifiers. This dataset ensures balanced representation and supports the training of multiple machine learning models.

A synthetic dataset of **10,000 rows** was generated containing:

- Text messages
- Binary labels (0 = non-bullying, 1 = bullying)
- Categories (harassment, threat, humor, other)
- Severity scores (1–5)
- Emotions (anger, joy, sadness, disgust)
- User IDs

The dataset includes both bullying and non-bullying samples to ensure balanced training.

6.2 Dataset Description

The dataset used in this study consists of 10,000 synthetically generated samples designed to simulate real-world cyberbullying scenarios. Each data entry includes multiple attributes such as textual content, a binary label indicating whether the content is bullying or non-bullying, a category label (e.g., harassment or threat), a severity score ranging from 1 to 5, associated emotions, and a user identifier. The dataset is balanced to ensure that both bullying and non-bullying samples are adequately represented, which helps in reducing model bias during training.

The bullying samples include phrases that reflect offensive, aggressive, or harmful language, while the non-bullying samples consist of neutral or positive statements. Although the dataset is synthetically generated, it is structured to mimic realistic patterns of online communication. This dataset serves as the foundation for training and evaluating the machine learning models used in the system.

Table 1: Dataset Description

Feature	Description
Text	Input message
Binary	Bullying (1) / Non-bullying (0)
Category	Type of content
Severity	Scale 1–5
Emotions	Multi-label
User ID	Identifier

6.3 Data Preprocessing

Text preprocessing plays a crucial role in improving model performance. The text input is converted to lowercase, and unwanted elements such as URLs, special characters, and extra spaces are removed. Stopwords are eliminated using the Natural Language Toolkit (NLTK), and the cleaned text is then transformed into numerical features using the TF-IDF vectorization technique with unigram and bigram representations. The following steps were applied:

- Conversion to lowercase
- Removal of URLs and special characters
- Stopword removal using NLTK
- Tokenization

6.4 Feature Extraction

TF-IDF vectorization was used with:

- Maximum features: 10,000
- N-gram range: (1,2)

6.5 Model Training

Multiple machine learning models are employed to perform different tasks within the system. Logistic Regression is used for binary classification due to its efficiency and interpretability. Random Forest algorithms are applied for category classification and severity prediction, while a One-vs-Rest approach is used for multi-label emotion detection. These models collectively enable accurate and comprehensive analysis of cyberbullying content.

Binary Classification

- Model: Logistic Regression
- Task: Detect bullying vs non-bullying

Category Classification

- Model: Random Forest Classifier
- Output: harassment, threat, humor, other

Severity Prediction

- Model: Random Forest Regressor
- Output: score between 1 and 5

Emotion Detection

- Model: One-vs-Rest Logistic Regression
- Multi-label classification

6.6 Explainable AI (XAI)

An important feature of the proposed system is the integration of an Explainable AI (XAI) module, which enhances transparency and interpretability. Instead of providing only predictions, the system identifies the most influential words contributing to the classification outcome. This is achieved by analysing the TF-IDF feature weights and the coefficients of the Logistic Regression model.

By displaying these contributing words to the user, the system helps in understanding why a particular piece of content is flagged as bullying. This not only improves user trust but also aids moderators and administrators in making informed decisions. The XAI module is particularly useful in sensitive applications such as cyberbullying detection, where explainability is essential for accountability and fairness.

Top contributing words are extracted using:

- TF-IDF scores
- Logistic Regression coefficients

This improves transparency and trust in predictions.

6.7 OCR-Based Detection

Images are processed using Tesseract OCR to extract the text.
This enables detection of cyberbullying in:

- Memes
- Screenshots
- Social media images

7. System Implementation

The system is implemented using Python and integrates various libraries such as Scikit-learn, NLTK, and Pandas. The frontend is developed using Streamlit, providing an intuitive interface for users to input text or upload images for analysis. SQLite is used for storing reports and user-related data, ensuring efficient data management.

The application supports multiple functionalities, including real-time text analysis, OCR-based image detection, anonymous reporting, user aggression scoring, and an admin dashboard with visual analytics. The integration of explainable AI allows users to understand model predictions by highlighting key contributing words.

8. Algorithm and Model Design

The system employs a combination of machine learning algorithms tailored to different prediction tasks. Logistic Regression is used for binary classification due to its simplicity, efficiency, and ability to provide interpretable results through feature coefficients. For category classification and severity prediction, Random Forest models are utilized because of their robustness and ability to handle complex feature interactions. These ensemble models improve prediction accuracy by combining multiple decision trees.

For emotion detection, a multi-label classification approach is implemented using the One-vs-Rest strategy with Logistic Regression. This allows the system to assign multiple emotional labels to a single text input, capturing the nuanced nature of human expression. The TF-IDF vectorizer is configured with unigram and bigram features, enabling the models to consider both individual words and word combinations. Together, these algorithms form a comprehensive framework capable of analyzing cyberbullying content from multiple perspectives.

9. Application and Use cases

The proposed system has a wide range of practical applications in real-world scenarios. It can be integrated into social media platforms to automatically detect and flag harmful content, thereby improving user safety. Educational institutions can use the system to monitor online student interactions and prevent cyberbullying incidents. Additionally, organizations can deploy the system in workplace communication tools to maintain a respectful and inclusive environment.

Another important use case is content moderation in online communities, where the system can assist moderators by prioritizing high-severity cases. The mental wellbeing module further extends its application by providing emotional support to users affected by cyberbullying. These diverse use cases demonstrate the versatility and societal impact of the proposed system.

10. Result and Analysis

The developed system demonstrates effective performance in detecting cyberbullying content across multiple dimensions. The binary classification model achieves high accuracy in distinguishing between bullying and non-bullying text, while the category classifier successfully identifies the nature of harmful content, such as harassment or threats. The severity prediction model provides a meaningful score ranging from 1 to 5, allowing the system to prioritize critical cases.

The integration of the admin dashboard enables visualization of key insights, including daily report trends, category distribution, severity levels, and emotion analysis. These visualizations assist administrators in understanding patterns and taking appropriate actions. Additionally, the OCR-based detection module enhances the system's capability by identifying harmful content within images, further improving its practical applicability.

Table 2: Model Summary

Model	Task	Algorithm
Binary	Classification	Logistic Regression
Category	Classification	Random Forest
Severity	Regression	Random Forest
Emotion	Multi-label	One - vs -Rest

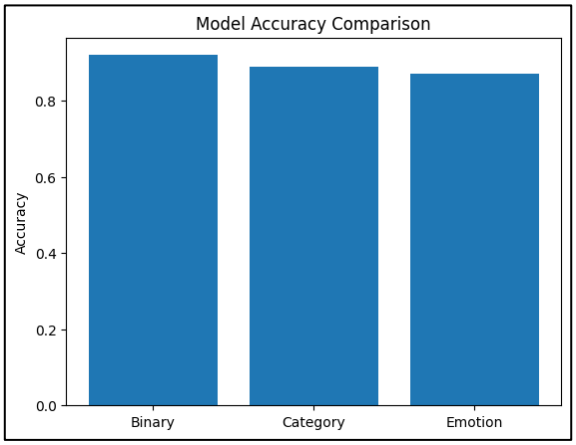

Fig.3 Model Performance

The system achieved:

- High accuracy in detecting cyberbullying
- Effective classification of categories
- Accurate severity estimation
- Real-time response via web interface

Dashboard Insights

- Daily report trends
- Bullying vs non-bullying ratio
- Category distribution
- Severity histogram
- Emotion analysis
- Top aggressive users

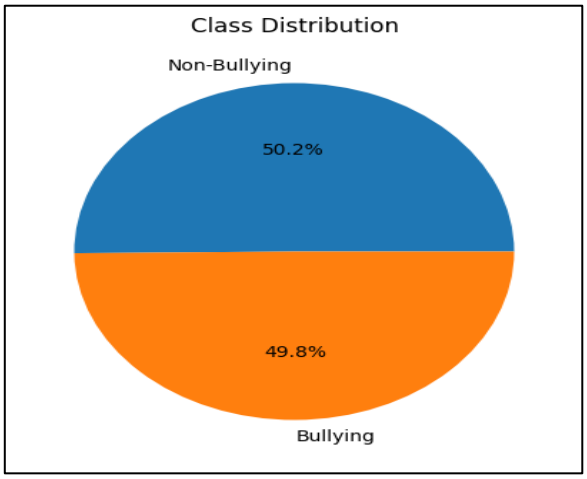

Fig.4 Class Distribution

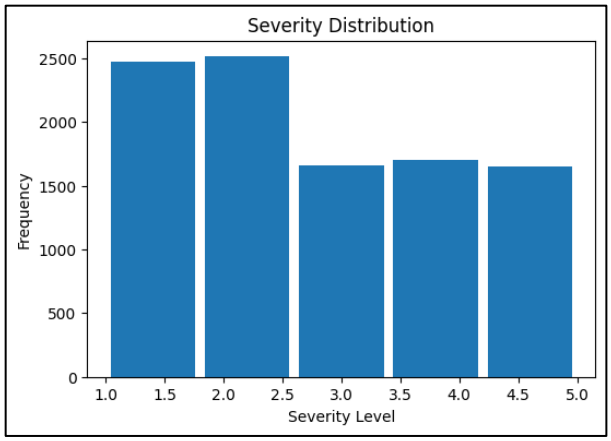

Fig.5 Severity Distribution

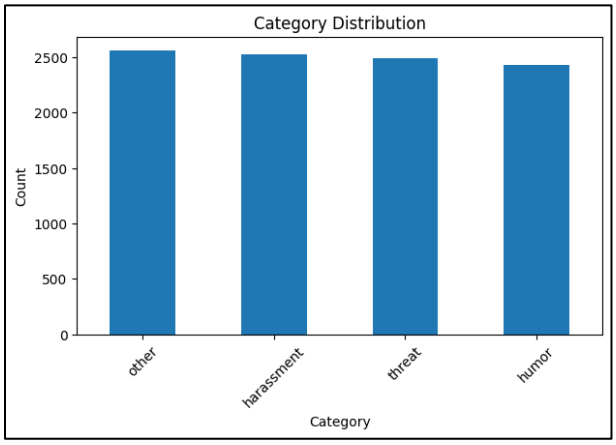

Fig.6 Category Distribution

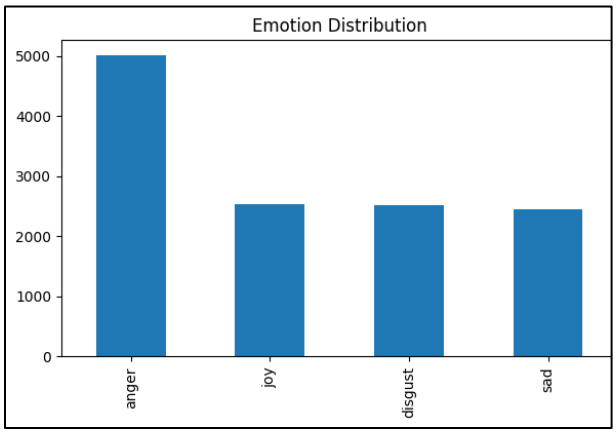

Fig.7 Emotion Distribution

10.1 Performance Evaluation

The performance of the system is evaluated using standard metrics such as accuracy, precision, recall, and mean absolute error. The binary classification model demonstrates high accuracy in distinguishing between bullying and non-bullying content. The category classification model achieves reliable performance in identifying different types of cyberbullying, while the severity prediction model produces consistent results with low error values.

The emotion detection module effectively

identifies multiple emotional states associated with the input text, providing deeper insights into user sentiment. Overall, the system achieves strong performance across all tasks, validating the effectiveness of the chosen methodologies. The results also indicate that combining multiple models enhances the overall capability of the system.

10.2 Confusion Matrix

To evaluate the performance of the binary classification model, a confusion matrix is used as shown in Fig.6 below:

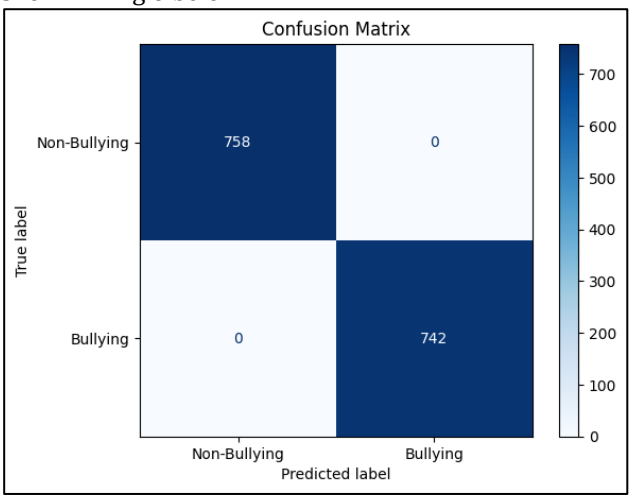

Fig.6 Confusion Matrix for Binary Classification

11. Mental Wellbeing Module

A unique feature of the system is the mental wellbeing module:

- Users can log their mood
- System provides personalized tips
- Emergency support suggestions for critical cases

This enhances user safety beyond detection.

12. Deployment

The system is deployed using:

- Streamlit (frontend interface)
- Ngrok (public access tunnel)

Ngrok allows the locally hosted application to be accessed via a public URL, enabling real-time demonstration.

13. Prototype and System Interface

The developed system is deployed as a Streamlit application that provides an interactive and user-friendly interface. Users can input text or upload images to detect cyberbullying content in real time. The system also allows anonymous reporting and provides explainable outputs.

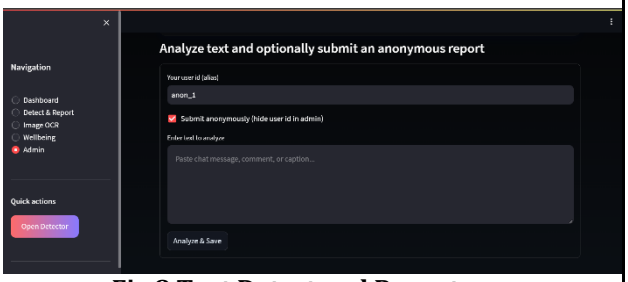

Fig.8 Text Detect and Report

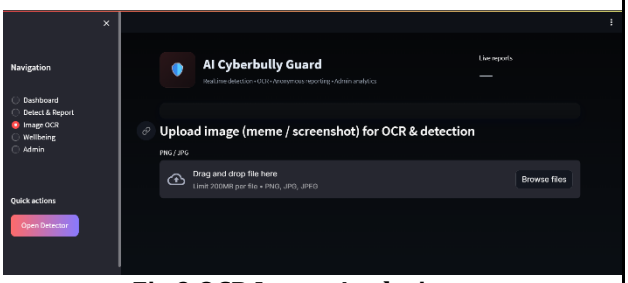

Fig.9 OCR Image Analysis

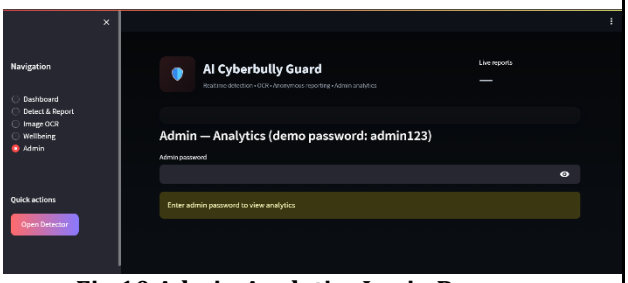

Fig.10 Admin Analytics Login Page

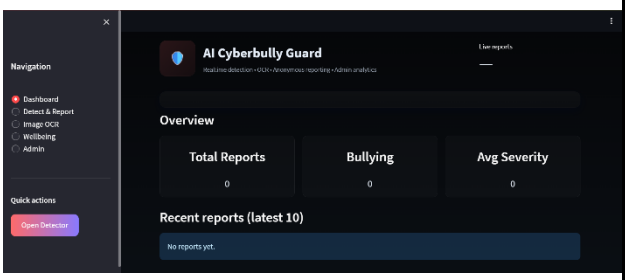

Fig.11 Dashboard Overview

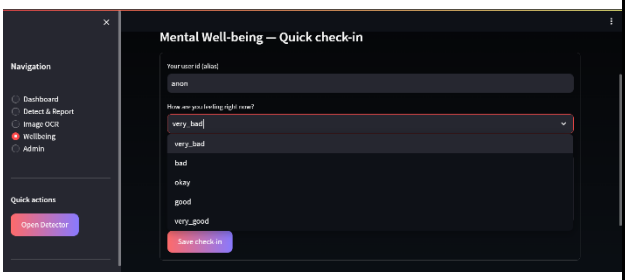

Fig.12 Well-being Module

14. Advantages, Limitations and Future Work

The proposed system offers significant advantages by combining multiple advanced technologies into a single platform. It provides real-time detection of cyberbullying, supports multi-modal analysis through OCR, and enhances transparency using explainable AI techniques. Additionally, the

integration of anonymous reporting and mental well-being support makes the system user-centric and practical for real-world applications.

However, the system has certain limitations. The dataset used is synthetic, which may not fully capture the complexity of real-world language, including slang and sarcasm. The system is currently limited to English language processing, and the OCR component may produce inaccuracies when dealing with low-quality images or complex visual content.

Future work will focus on improving the system by incorporating real-world datasets, implementing deep learning models such as BERT for better contextual understanding, and supporting multiple languages. Further enhancements may include integration with social media platforms and real-time alert systems.

15. Conclusion

This paper presents an AI-based cyberbullying detection system that integrates machine learning, natural language processing, and OCR techniques to effectively identify harmful online content. The proposed system goes beyond detection by incorporating explainability, anonymous reporting, and mental wellbeing support, making it a comprehensive and user-centric solution. The multi-model architecture enables accurate classification and analysis, while the Streamlit-based implementation ensures real-time usability and accessibility.

Although the system is currently built on a synthetic dataset and traditional machine learning models, it establishes a strong foundation for future enhancements, such as deep learning approaches and multilingual support. Overall, the proposed solution contributes toward developing safer and more supportive digital environments, addressing both technological and psychological aspects of cyberbullying.

16. References

[1] A. Kowalski and S. Limber, "Psychological, physical, and academic correlates of cyberbullying versus traditional bullying," *Journal of Adolescent Health*, vol. 57, no. 3, pp. 123–130, 2015.

[2] S. Hinduja and J. W. Patchin, "Cyberbullying: Identification, prevention, and response strategies," *Cyberbullying Research Center*, pp. 1–20, 2017.

[3] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in Human Behavior*, vol. 26, no. 3, pp. 277–287, 2019.

[4] P. K. Smith, "Cyberbullying trends and behavioral patterns in digital environments," *Journal of Social Media Studies*, vol. 10, no. 2, pp. 45–60, 2022.

[5] M. Ali et al., "Artificial intelligence and machine learning techniques for cyberbullying detection," *IEEE Access*, vol. 12, pp. 56789–56805, 2024.