Abstract

Cyberbullying and online hate speech have become significant social challenges in modern digital communication, especially on social media platforms, messaging apps, and online communities. Harmful messages can negatively affect individuals' mental health, self-esteem, and safety, making it essential to detect and prevent abusive content before it spreads. This project focuses on developing a Cyberbullying and Hate Speech Detection System using machine learning-based toxicity classification to help identify and monitor harmful text in real time.

The system is built using HTML, CSS, and JavaScript for the user interface and integrates the TensorFlow.js Toxicity Model to analyze user-entered text. The model detects six key categories of abusive behavior, including toxic language, severe toxicity, identity attacks, threats, insults, and obscene expressions. Once the input is analyzed, the system displays the prediction result indicating whether the comment is harmful or safe along with its confidence score. The performance of different toxicity categories is visually represented through bar and pie charts using Chart.js for better interpretation.

To enhance security and administrative control, the system features a login-based access for the admin, who can review harmful comments. Any comments that exceed a predefined severity level are blocked automatically and added to a flagged list, allowing the admin to take further action. A reporting mechanism with an audible alert provides an interactive experience for immediate attention. Additional features such as theme switching, comment history storage, and a clean user interface improve overall usability.

This project supports safer online communication by helping users become aware of cyberbullying while assisting moderators in monitoring harmful content more efficiently.