Title of the Project : BullyBlock AI: A Real-Time AI-Based Cyberbullying Detector and Reporter

Name of the Students : S.ABINAYA

J.DEEPIKA

Register Number : 211423104005

211423104110

Name of the Guide : Mrs.CINTHUJA

ABSTRACT

Cyberbullying has become a major concern in the digital era, especially among students and young people, leading to anxiety, depression, and poor academic performance. Existing AI tools typically stop at detecting harmful content, requiring victims to report abuse manually. To address these gaps, BullyBlock AI provides an end-to-end solution that not only detects cyberbullying but also evaluates its severity and automates the reporting process. The proposed system processes online communications using advanced Natural Language Processing (NLP) techniques and pre-trained transformer models such as BERT, integrated with the spaCy framework for efficient text parsing. A custom classifier categorizes the intensity of abuse into Low, Medium, or High levels. When severe or repeated harassment is identified, an automated alert is triggered through SMTP to notify authorities and support teams in real time. Beyond detection, BullyBlock AI incorporates sentiment analysis and interactive data visualization to monitor evolving trends in user interactions. Positive and negative sentiments of flagged messages are tracked over time, offering stakeholders actionable insights into the overall communication climate. The modular design supports easy deployment on web and mobile platforms, making it adaptable for schools, social media monitoring teams, and law-enforcement agencies. This smart, proactive system not only provides early warning and rapid intervention but also establishes a scalable framework for safer digital communities. To ensure reliability and fairness, BullyBlock AI incorporates a robust training and evaluation pipeline. Diverse, publicly available cyberbullying datasets are preprocessed for balanced class distribution and subjected to cross-validation to minimize bias. The model is continuously fine-tuned with fresh data so it can recognize evolving slang, code-switching, and multilingual expressions common in online abuse. The system achieves an overall accuracy of 0.88 (88%).