# GENAI ASSIGNMENT 2

NAME: M NIRANJAN

CLASS:6E

SRN:PES2UG23CS308

TOPIC: CyberBullying Detector

The rapid expansion of social media and online discussion forums has inadvertently created fertile ground for cyberbullying, hate speech, and toxic behavior. Manual moderation of these platforms is resource-intensive and often unable to keep pace with the volume of user-generated content, leading to delayed responses and unsafe online environments. This project proposes an automated **Cyberbullying Detection System** utilizing **Deep Learning** and **Natural Language Processing (NLP)** to identify and flag toxic comments in real-time.

**Methodology** The system is built upon the **BERT (Bidirectional Encoder Representations from Transformers)** architecture, specifically utilizing the unitary/toxic-bert model. This model has been fine-tuned on the Jigsaw Toxic Comment Classification Challenge dataset. It employs a multi-label classification approach to categorize input text into six distinct classes of toxicity: *toxic, severe toxic, obscene, threat, insult,* and *identity hate*.

**Implementation** The solution is implemented using Python within the **Google Colab** environment, leveraging the Hugging Face transformers library for efficient model inference. For demonstration and testing purposes, an interactive web interface is integrated using the **Gradio** library, allowing for real-time text analysis.

**Significance** The system demonstrates high efficacy in parsing semantic context to distinguish between benign and harmful text. By automating the detection process, this tool aims to assist platform administrators in filtering harmful content efficiently, reducing the psychological impact of cyberbullying, and fostering healthier online communities.

# OUTPUT SCREENSHOTS