**ABSTRACT**

The rapid growth of social media has led to increased instances of toxic behavior, including hate speech, insults, and passive-aggressive language. Traditional toxicity detection models often struggle to identify context-dependent toxicity, such as sarcasm or indirect insults. This project presents a **context-aware toxicity detection system** that leverages advanced Natural Language Processing (NLP) techniques and transformer-based models, specifically fine-tuned versions of BERT and RoBERTa. The system preprocesses user-generated content, tokenizes the text, and performs multi-label classification to detect various forms of toxicity, including toxic, insult, threat, obscene, and identity-hate categories. A threshold-based mechanism ensures only high-confidence toxic predictions are flagged, reducing false positives. By focusing on contextual nuances and leveraging the power of deep learning, the system improves accuracy in distinguishing between harmful and non-harmful content. This project contributes to safer online environments by providing a tool for real-time moderation and intelligent filtering of offensive language on digital platforms.