Abstract

Social media has revolutionized the way people communicate and share opinions, offering platforms that are open, fast, and global. However, this openness also exposes users to toxic behavior such as hate speech, cyberbullying, harassment, and offensive language. Detecting such harmful content is essential to maintaining a safe online environment. Traditional toxicity detection systems primarily focus on keyword-based or rule-based methods, which often ignore the broader context in which a message is delivered. These methods may misclassify non-toxic comments as toxic or fail to detect toxicity masked by sarcasm, indirect references, or coded language.

This mini project aims to develop a **context-aware toxicity detection model** that leverages modern Natural Language Processing (NLP) techniques to analyze content within its conversational or semantic context. By incorporating context such as previous comments in a thread, user behavior patterns, and conversational tone, the system can better understand the intent behind a message. Advanced models like BERT (Bidirectional Encoder Representations from Transformers) are utilized to capture nuanced meanings and relationships between words in a given context.

The project involves collecting and preprocessing real-world social media datasets, training machine learning models on annotated toxic and non-toxic examples, and evaluating performance using metrics such as accuracy, precision, recall, and F1-score. Results show that context-aware approaches significantly outperform traditional models, particularly in cases involving implicit or indirect toxicity. This work highlights the importance of contextual understanding in developing smarter, more reliable content moderation tools to foster healthier online communities.