Abstract

We present a prototype **context-aware toxicity detection** system for social media comments that combines multiple analysis layers to identify both overt and subtle forms of harmful language. First, the system leverages pre-trained transformer pipelines to perform **sentiment analysis** and **sarcasm detection**, enabling it to distinguish literal negativity from ironic or sarcastic expressions. Next, a simple **conversational context tracker** accumulates prior comments to provide dialogue awareness—laying the groundwork for future extensions that incorporate thread-level cues. A **rule-based filter** then applies heuristic checks (e.g., flagging comments that are simultaneously ironic and negative) to catch indirect insults that may evade traditional keyword approaches. Finally, a fine-tuned BERT classifier performs direct toxicity classification on the comment text, with adjustable confidence thresholds to balance precision and recall. Through this multi-layered pipeline—combining NLP-based insight, lightweight context modeling, and robust classification—the framework aims to reduce false negatives in detecting masked toxicity (such as sarcasm plus insult) without compromising on true positives for overtly abusive language. This modular design can be expanded with richer context features, more sophisticated rules, and advanced dialogue-aware models to enhance reliability and scalability in real-world social media moderation.