ABSTRACT

With the exponential growth of social media and digital platforms, the detection of hate speech has become a critical task for ensuring safe online interactions. Conventional hate speech detection systems often rely on lexical and surface-level semantic cues, which limits their ability to recognize subtle and context-dependent forms of abuse such as sarcasm, disguised toxicity, or emotionally layered expressions. To address this limitation, we propose an **emotion-calibrated hate detection framework** that integrates artificial empathy into classification pipelines by leveraging emotion signals as auxiliary features. The framework is designed as a two-stage pipeline: in the first stage, emotion classification models are trained on the GoEmotions dataset using multiple architectures including BERT, DistilBERT, LightGBM, and FastText; in the second stage, the predicted emotion logits and labels are incorporated into hate speech classification on the Jigsaw Toxic Comment dataset. This design enables the system to capture not only overt hate but also covert and emotionally nuanced toxicity that traditional systems tend to overlook. Experimental results demonstrate the effectiveness of the proposed approach, where the FastText emotion classifier achieved an accuracy of **60.33%** across 28 emotions, while the DistilRoBERTa-based hate classifier reached an accuracy of **96.16%** with a macro F1-score of **0.88**. Furthermore, the combination of BERT emotion features with LightGBM yielded strong interpretability by highlighting which emotions most strongly influenced classification decisions. These results confirm that embedding emotional intelligence into automated hate detection systems significantly improves robustness, interpretability, and fairness. Overall, the study lays the groundwork for developing **emotion-aware, socially responsible AI models** that go beyond punitive moderation to create empathetic, transparent, and human-aligned solutions for online safety.