TikCare: A Novel DOM-Integrated Pipeline for the Real-Time Detection and Obfuscation of Harm Proliferation in Algorithmic Video Feeds

Behairy, Sohibe (School: Obour STEM School)
Hassona, Youssef (School: Obour STEM School)

Preadolescent exposure to harmful short-form video (SFV) content now exceeds 60% (eSafety Commissioner). Current moderation pipelines, particularly TikTok's "Filter & Refine" architecture, leak implicit harm by escalating only flagged candidates to high-fidelity rankers, leaving subtle threats undetected at the routing stage. Existing multimodal classifiers use late-fusion architectures processing visual and textual streams independently, failing to detect cross-modal inconsistencies such as benign visuals paired with harmful overlays or algospeak, where stream semantics contradict overall intent. Hashtag-driven datasets miss low-virality long-tail content within filter bubbles. TikCare addresses these gaps via a DOM-level, pre-render interception layer with a two-tier safety cascade. Tier 1 deploys a fine-tuned EfficientNetV2 to reject explicit harm; all remaining content undergoes mandatory Tier 2 validation by Gemini 2.5 Flash, processing joint visual-text sequences via cross-modal fusion in a unified embedding space, resolving the late-fusion gap and enabling detection of cross-modal contradictions. Trained on a curated Egyptian-Arabic "Active Engagement" dataset (N=1,583), few-shot prompts encode cultural-linguistic nuances for Egyptian digital contexts. Evaluated on a stratified held-out test set (N=491; 31% of corpus), TikCare achieved F1=92.2% and Recall=96.0%, minimizing false negatives critical in child-safety contexts, surpassing MTikGuard (F1=89.45%). The pipeline uses background prefetching, processing five videos within 0.3–3.5s against a mean duration of 42.7s, ensuring harmful content is blocked before the user scrolls. TikCare delivers a scalable, culture-aware framework for real-time harmful content enforcement across diverse digital environments.

Awards Won: