ICCP/SL/OP/349

Enhancing child safety online: A multi-modal not safe for work content detection system

Erandaka ER*, Sandunwala ST, Madurapperuma MTSK, Rathnayaka RMHI

[1]*Department of Statistics and Computer Science, University of Kelaniya, Sri Lanka*
*erandak-ec20037@stu.kln.ac.lk

Background: The proliferation of digital platforms has significantly increased children's exposure to Not Safe for Work (NSFW) content, encompassing adult and violent material, posing substantial risks globally. Current video detection systems have limitations in automation, real-time processing, and comprehensive detection of violent content. Furthermore, existing models do not concurrently identify multiple NSFW content types (including text and imagery), highlighting a critical research gap. This study was conducted to propose a multi-modal NSFW detection system designed to surpass existing accuracy benchmarks to safeguard children in online environments.

Method: The system employed a fine-tuned DistilBERT model for NSFW text detection and a YOLOv11n model for image and video analysis. An automated OpenCV pipeline extracts video frames at 15 FPS, facilitating real-time processing. The training process utilizes a custom dataset alongside publicly available repositories from Hugging Face and Kaggle. The detection models were integrated into a web-based browser application through a streamlined data pipeline.

Results: The system achieved 91% accuracy in text detection with DistilBERT and 80% accuracy in image and video analysis using YOLOv11n, trained on a dataset comprising over 15,000 annotated samples. Integration into a functional web-based application demonstrated effective NSFW content moderation capabilities.

Conclusions: This research introduces a novel multi-modal NSFW detection framework capable of real-time content moderation, significantly contributing to enhanced child safety in digital spaces within Sri Lanka. Future recommendations include expanding dataset diversity, optimizing model architectures, and implementing the system in policy-driven environments, such as educational institutions and national safety networks.

Keywords: Not safe for work, child protection, machine learning, DistilBERT, YOLO11n.