IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

Digital Purifier

Shaima Shahul[1], Renjini L R[2], Harikrishnan S R[3]
Student, MCA, CHMM College for Advanced Studies, Trivandrum, India [1]
Associate Professor, MCA, CHMM College for Advanced Studies, Trivandrum, India [2]
Associate Professor, MCA, CHMM College for Advanced Studies, Trivandrum, India [3]

Abstract: *The digital age is increasingly challenged by the spread of manipulated media (deep fakes) and the pervasive issue of online hate speech and toxic comments. This project proposes an integrated system, deployed as a Flask web application, to address these critical issues. The system will comprise two primary modules: a deep fake detection module and a toxic comment analysis and removal module. The deepfake detection module will employ Generative Adversarial Networks (GANs) to accurately classify uploaded images as either fake or real, analyzing subtle inconsistencies indicative of manipulation. The toxic comment module will utilize BERT (Bidirectional Encoder Representations from Transformers) to identify and categorize hate speech and toxic language within user-generated text. Upon detection of toxic comments, the system will automatically remove the offending content. Furthermore, all detected toxic comments, along with associated metadata, will be systematically recorded and saved to an Excel spreadsheet for future analysis and moderation purposes. This comprehensive platform aims to enhance online safety and information integrity by providing real-time detection and mitigation of both deepfakes and toxic content, while maintaining a detailed record for ongoing review and improvement*

Keywords: GAN, BERT, NLP, Pytorch, Machine Learning, Deep Learning

DOI: 10.48175/568