Comprehensive Approach to Detecting Harmful Content in Digital Media

TEAM MEMBERS: S RAMYA 22AM042

Abstract

With the increasing prevalence of social media, harmful content such as hate speech, cyberbullying, and misinformation has become a significant concern. This project aims to develop an automated system for detecting harmful content in social media posts.

Problem Statement

To build a system for detecting harmful contenSt on social media, this project focuses on leveraging machine learning and natural language processing (NLP) techniques to classify posts as harmful or safe. The goal is to develop a scalable, real-time detection solution to enhance online safety by identifying hate speech, cyberbullying, and misinformation.

Architecture Flow

Given a content, the system operates as follows:

1. Text:
    - Checks if the content is in English. If not, it translates the text to English using the Google Translate API.
    - The translated text is then fed into the harmful content detection model.
2. Image:
    - The image is converted into a description using an image-to-text processing tool.
    - This description is then processed as in step 1 for harmful content detection.
3. Video:
    - Extracts images from the video.
    - Each extracted image is processed as in step 2 to detect harmful content.

GUIDE:

Dr. R. Gowri

Head Of Department (CSE(AIML))