AI-Powered Detection of Online Harms

Team Details

Team Name:

Team Leader: Jagabandhu Prusty

Problem Statement

With the rapid growth of social media and online communication, harmful content such as cyberbullying, hate speech, and misinformation is spreading at an alarming rate. Manually moderating such content is inefficient and unreliable. The need for an automated, AI-driven solution is critical to ensure a safer online environment.

Solution Overview

The AI-powered detection of online harms is a web-based platform that allows users to upload content for analysis without requiring sign-in or login. The system leverages Googles AI models to detect harmful content and provide actionable insights. The solution ensures quick and accurate moderation of online content, helping users and organizations maintain a healthy digital space.

Unique Selling Proposition (USP)

- **No login required:** Users can analyze content anonymously.
- **Real-time detection:** AI models analyze text, images, and videos instantly.
- **High accuracy:** Google AI models and NLP techniques improve detection efficiency.
- **Scalable API support:** Can be integrated into existing platforms for automated moderation.
- **User-friendly interface:** Simple and intuitive web design for seamless experience.
- **Sentence Replacement:** Provides alternative suggestions for harmful sentences.
- **Control Feature:** Users can enable/disable automatic moderation through a control button.

Features

- Upload content (text, images, or videos) for analysis
- AI-based detection of harmful content
- Categorization of detected harms (e.g., hate speech, cyberbullying, misinformation)
- API integration for third-party platforms
- Real-time moderation reports and suggestions
- **Google Video Intelligence API:** Processes and analyzes video content.

- **Google Transcription API:** Transcribes and detects harmful speech in audio/video.
- **Sentence Editing Feature:** Users can edit detected harmful content and receive AI-generated alternate suggestions.
- **Control Button:** Allows users to enable or disable automatic moderation.

Process Flow

1. **User uploads content** (text, image, or video).
2. **AI model processes content** using NLP, Video Intelligence API, and Transcription API.
3. **System detects harmful content** and categorizes it.
4. **User receives a report** with the results and suggested actions.
5. **Optional API integration** for external platforms to automate moderation.
6. **Sentence Replacement Feature:** Users can modify flagged sentences with AI-generated suggestions.

Architecture Diagram

User Upload Content AI Model Processing Harm Classification Report Generation Sentence Replacement

Technologies Used

- **Frontend:** HTML, CSS, JavaScript
- **Backend:** PHP, MySQL
- **AI Models:** Google Gemini API, NLP, Image Recognition
- **APIs:** Google Video Intelligence API, Google Transcription API
- **Hosting:** InfinityFree (for initial deployment)
- **API Integration:** Custom-built API for content scanning

Future Enhancements

- **Improved AI Accuracy:** Fine-tune models with more datasets.
- **Expanded Language Support:** Multilingual detection.
- **Advanced Image & Video Analysis:** Detect harmful memes and deepfake content.
- **Browser Extension:** Real-time scanning while browsing.
- **Mobile App Development:** Extend functionalities to mobile users.
- **Advanced Sentence Editing Tool:** Provide contextual suggestions with AI-powered grammar checks.

- **Dashboard for Users:** View history of flagged content and changes made.

Additional Details

- **Demo Link:** [To be added]
- **Prototype Hosted At:** Not required for now

This document provides a comprehensive overview of the AI-powered online harm detection project. The platform aims to enhance digital safety by leveraging AI-driven content moderation.