CloutCheck - AI-Powered Reputation & Content Safety Scanner for Influencers & Brands

Project Overview

CloutCheck is a SaaS-based AI platform designed to scan public social media content of influencers, job candidates, and public figures across platforms (YouTube, TikTok, Twitter, Instagram) to generate a *Reputation Risk Report*. It flags harmful or sensitive content (hate speech, drugs, sexual innuendos, offensive imagery, religious extremism, etc.) using multimodal AI (vision, speech, text).

Problem Statement

In the age of cancel culture, one viral post can damage a brand. HR teams, brands, PR firms, and concerned parents have no scalable way to audit a person's online content for reputation risks. Manual checks are inefficient and incomplete.

Solution Summary

ReputeCheck automatically:

- Scrapes social media content based on public handles
- Analyzes videos, text, speech, and images
- Flags harmful content categories
- Assigns a Reputation Score
- Generates downloadable reports

Target Users

- Brand managers & influencer agencies
- HR teams for background checks
- NGOs and legal professionals
- Parents/guardians
- Compliance & audit departments

Core Features (MVP)

Feature	AI Component
Video Scanning	YOLOv8 / ResNet for object & scene detection
Speech Analysis	Whisper for speech-to-text

Text/NLP	RoBERTa/BERT for hate speech, sexual content etc.
Social Scraping	APIs for YouTube, Twitter, TikTok, Instagram
Dashboard	For scan management, report generation, user controls

Research Angle

Research Problem:
“Use multimodal deep learning (vision + speech + NLP) to build a *Reputation Risk Classifier* for public digital content.”
OR
“Multimodal Reputation Risk Classification via Joint Learning from Vision, Speech, and Text Modalities”

Design a unified **multimodal AI model** that combines:

- Visual detection (nudity, weapons, drugs)
- Speech/NLP classification (hate, offensive language)
- Textual content sentiment & toxicity analysis

Possible contributions:

- Fusion architecture (early, late, or attention-guided)
- Novel benchmark for reputation risk
- Bias/fairness evaluation in content moderation AI
- Dataset creation & annotation

Evaluation Metrics

- Accuracy / Precision / Recall per category (hate, sexual)
- Confusion matrix of misclassified risk types
- Reproducible benchmark performance
- User study for usefulness (optional)

Who Needs This?

- Brand managers evaluating influencers
- PR agencies before campaigns
- NGOs or legal teams checking for hate speech or fake news

Project Plan (Timeline)

Semester 7:

- Literature review & dataset creation
- Build scraping pipeline & MVP UI
- Train baseline models
- Start writing paper draft

Semester 8:

- Final model fusion & experimentation
- Full system integration (dashboard + API)
- Paper submission to conference/workshop
- SaaS deployment MVP

AI/ML Tech Stack

- YOLOv8 / ResNet (visual moderation)
- Whisper (speech-to-text)
- BERT/RoBERTa (hate speech, etc.)
- Transformers for fusion
- FastAPI (backend), MongoDB, React (frontend)

Datasets

- YouTube8M
- HateXplain (text)
- https://www.kaggle.com/datasets/khushhalreddy/violence-detection-dataset (perfect for us)
- https://www.kaggle.com/datasets/wasifullahcs/tiktok-video-dataset
- https://www.kaggle.com/datasets/fangfangz/audio-based-violence-detection-dataset
- https://github.com/EBazarov/nsfw_data_source_urls
- https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset
- https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset
- https://github.com/hate-alert/HateMM
- https://www.kaggle.com/datasets/wajidhassanmoosa/multilingual-hatespeech-dataset
- https://hasocfire.github.io/hasoc/2024/dataset.html
- https://www.kaggle.com/datasets/aadyasingh55/sexism-detection-in-english-texts
- https://www.kaggle.com/datasets/victorcallejasf/multimodal-hate-speech
- https://www.kaggle.com/datasets/adilshamim8/toxicity-detection-context

💼 Monetization Plan

- Freemium: Limited scans/month
- Pro Tier: Unlimited scans, API access, branded reports
- Custom plans for HR or influencer firms

🕹️ Output Deliverables

- SaaS product demo (web-based dashboard)
- Published research paper (top-tier AI workshop)
- GitHub repo with open code (and optionally dataset)
- Risk reports with score explanation

🌟 Summary Pitch

> ReputeCheck is an AI-powered tool that helps brands, HR teams, and parents assess public social media content for reputational risks. Using cutting-edge multimodal AI, it flags hate speech, drug glorification, sexual content, and more — generating a reputation score and actionable report in minutes.

Ready to make the internet safer, smarter, and reputation-aware!

Literature review and additional docs:

https://manojbhor.medium.com/nsfw-classifier-nudity-classification-on-mobile-and-edge-devices-bae14c171a89
https://www.researchgate.net/publication/337273753_RWF-2000_An_Open_Large_Scale_Video_Database_for_Violence_Detection
https://arxiv.org/abs/2305.03915
https://www.researchgate.net/publication/370625458_Using_Deep_Learning_to_Detect_Islamophobia_on_Reddit
https://github.com/Slainteee/Jatayu-ContentModeration (perfect for us)