CloutCheck AI - Multimodal Reputation Risk Analysis Platform

Final Year Project Proposal

Session FALL 2025

A 4th Year Student

BS in Computer Science

Department of CS

Fast School of Computing

Fast National University, Karachi Campus

21 September 2025

Project Registration

Project ID (for office use)	
Type of project	[ ] Traditional [✓] Industrial [ ] Continuing
Nature of project	[ ] Development [✓] Research & Development [ ] Research
Sustainable Development Goals(SDGs)	[ ] Good Health and Well-Being [ ] Quality Education [✓] Industry, Innovation, and Infrastructure [ ] Gender Equality [✓] Decent Work and Economic Growth [ ] Climate Action
Area of specialization	[✓] Artificial Intelligence (AI) [✓] Data Science and Analytics [ ] Internet of Things (IoT) [ ] Blockchain [ ] Mobile App Development [✓] Web Development [ ] Cybersecurity [ ] Game Development [✓] Natural Language Processing (NLP) [ ] Other

Project Group Members

Sr.#	Reg. #	Student Name	CGPA	Email ID	Phone #	Signature
(i)	Group Leader	Syed Uzair Hussain	3.66	k224212@nu.edu.pk	0320 3016748	
(ii)	Group Member	Saad Ahmed	3.17	k224345@nu.edu.pk	0312 2260640	
(iii)	Group Member	Huzaifa Bin Khalid	2.24	k224223@nu.edu.pk	0313 3519640	

Declaration: FYP group members have cleared all prerequisite courses For FYP-I as per their degree requirements.

Project Abstract

In today's hyper-connected digital landscape, reputation has become one of the most valuable assets for individuals and organizations alike. For brands a single controversial post, video, or statement from their influencer partner can rapidly spiral into reputational damage, financial loss, and diminished trust. Existing tools for influencer vetting and social listening primarily focus on sentiment tracking and keyword monitoring, but they lack the capability to evaluate content across diverse modalities and consolidate findings into a unified risk score. This gap underscores the pressing need for an intelligent, automated, and multimodal reputation risk assessment platform.

CloutCheck AI is proposed as a cutting-edge solution to this challenge. The platform leverages artificial intelligence to analyze multimodal content, including text, images, videos, and audio, collected from various social media platforms and online sources. By employing advanced natural

language processing, computer vision, and speech recognition models, the system identifies harmful or controversial elements such as hate speech, extremism, nudity, substance abuse, or inappropriate political messaging. These findings are aggregated into a transparent, standardized risk score, enabling brands and decision-makers to evaluate influencer credibility objectively and efficiently.

The proposed system introduces a pipeline for content analysis and an interactive user interface. Together, they provide secure access, seamless data integration, real-time monitoring, and professional reporting tools. Beyond its academic contribution, CloutCheck AI offers direct industry impact by reducing manual screening efforts, ensuring consistency in evaluation, and supporting proactive risk management strategies. This innovative platform holds potential to redefine how organizations assess partnerships, safeguard brand identity, and navigate the complexities of the digital era.

Introduction

The proposed project, **CloutCheck AI: An AI-Powered Multimodal Reputation Risk Analysis Platform**, addresses the urgent need for comprehensive influencer and public-figure reputation assessment in the digital era. Social media platforms have transformed into powerful channels for branding and outreach, but they also carry reputational risks due to the vast and unregulated nature of online content. Traditional monitoring methods rely on manual vetting or text-focused sentiment tools, which are insufficient for detecting multimodal risks such as controversial images, harmful audio content, or context-driven video clips. To bridge this gap, CloutCheck AI will deliver an automated solution that integrates text, vision, and speech analysis into a unified risk scoring framework, thereby supporting proactive and evidence-based decision-making.

The main objectives of this project are threefold:

1. **Automated Multimodal Content Analysis** – Develop/ Finetune AI models capable of analyzing text, images, videos, and audio content from diverse online platforms.

2. **Reputation Risk Scoring** – Design a transparent scoring system that quantifies risk severity across predefined categories such as hate speech, extremism, nudity, substance abuse, and controversial politics.

3. **Decision Support & Reporting** – Build an interactive dashboard with visualization tools and professional reporting features to aid stakeholders in hiring, partnership, or compliance-related decisions.

The project will be structured into the following sub-tasks: content scraping and integration using APIs and custom scrapers, natural language processing for textual sentiment and toxicity detection, computer vision models for image and video frame classification, speech-to-text transcription combined with toxic language recognition, and multimodal data fusion for risk indexing. Techniques from deep learning, such as transformer-based NLP (e.g., BERT, RoBERTa), vision models (e.g., CLIP, ViT), and Whisper for speech analysis, will be employed.

Success will be measured by evaluating the system's accuracy in detecting high-risk content, consistency of risk scores across test datasets, and user feedback on dashboard usability. Evidence of success will include performance benchmarks, comparison with existing tools, and demonstration of a functional SaaS platform documented in the final dissertation.

Success Criterion

The success of this project will be defined by its ability to meet the stated objectives in a verifiable and measurable manner. A successful outcome will be demonstrated if the system can automatically extract, process, and analyze multimodal content; text, images, videos, and audio from selected social media platforms and generate a consolidated reputation risk score with reasonable accuracy and consistency.

At a minimum, the system should be able to:

- Integrate with at least two major platforms (e.g., X and Instagram) for profile content extraction.
- Accurately detect and classify high-risk content categories such as hate speech, nudity, violence, or drug-related material with benchmarked performance against test datasets.
- Produce an overall reputation index score that combines multimodal analysis into a transparent and interpretable framework.
- Provide an interactive dashboard that visualizes risk metrics and allows users to generate professional, branded reports.

These deliverables will be supported by quantitative evaluation (e.g., accuracy, precision, recall of detection models), qualitative feedback (user satisfaction with dashboard usability), and comparative analysis against existing reputation monitoring tools. The project will be considered successful if these minimum requirements are satisfied, ensuring that the core contribution of multimodal automated reputation scoring is achieved.

Beyond this baseline, the project may exceed expectations by offering additional capabilities such as real-time monitoring, explainable AI reasoning, or scalability across multiple cloud environments. Ultimately, project success will be validated through supervisor approval and, if applicable, positive reception by external evaluators or client stakeholders.

Related work

Reputation assessment and content moderation in social media draw on two closely related research streams:

1. studies that conceptualize influencer reputation and the social consequences of influencer behavior, and
2. technical work on automated content moderation and multimodal risk detection. Together, these streams provide the theoretical grounding and methodological approaches that inform the design of CloutCheck AI.

Influencer reputation and motivation

Several works establish the multi-dimensional nature of influencer reputation and document the negative consequences that reputation breaches can cause for brands and stakeholders. Ryu and Han provide a validated multidimensional scale for influencer reputation, identifying core constructs such as authenticity, expertise, communication, and influence; this framework helps operationalize what "reputation" means and guides the choice of risk categories in CloutCheck AI [1]. Complementing measurement research, recent conceptual reviews on the darker effects of influencer activity synthesize ethical, psychological, and societal harms arising from authenticity breaches and controversial behavior, strengthening the practical motivation for automated reputation monitoring [2]. In the proposal, these papers will be cited to justify the selection of reputation dimensions, define the high-level objectives of the platform, and motivate the inclusion of non-toxicity risk categories (e.g., authenticity/brand fit, political controversy).

Multimodal moderation and toxicity detection

The technical literature demonstrates that multimodal approaches (text + image + audio/video) substantially improve detection of harmful content relative to single-modality systems. Kim et al. (WSDM) show that fusing textual and visual signals improves influencer profiling for marketing applications and reduces false positives compared to text-only models; this supports our decision to adopt multimodal fusion for reputation scoring [3]. ToxVidLM (ACL 2024) presents an end-to-end multimodal architecture for toxicity detection in videos (combining speech, visual frames and text), which is directly relevant to CloutCheck's audio/video analysis pipeline and provides architectural and evaluation guidance [4]. The MuTox dataset and methods highlight the importance of multilingual, audio-centric toxicity resources — an important consideration for CloutCheck when handling non-English or code-mixed content in audio channels [5]. In the review, these works will be

referenced for model architectures, modality-specific preprocessing (e.g., speech transcription via Whisper-like pipelines), and evaluation metrics for toxicity detection.

Handling modality asymmetry and missing data

Real-world social data are heterogeneous: many posts lack certain modalities (e.g., image-only posts, text-only captions). Recent methodological contributions address asymmetric modality importance and missing-modality inference. The AM3 (Asymmetric Mixed-Modality) perspective analyzes how modality imbalance affects moderation outcomes and suggests fusion strategies that weight stronger modalities more heavily [6]. The Multimodal Guidance Network offers techniques to infer missing modalities or to compensate for absent signals during inference, which directly informs CloutCheck's robustness strategy when some content modalities are unavailable [7]. These papers will be cited to justify the fusion choices (early, late, and attention-guided fusion) and to motivate fallback inference mechanisms.

Surveys and large-model perspectives

A contemporary survey of hate speech moderation emphasizes the shift toward multimodality and the growing role of large language and multimodal models for moderation tasks [8]. This survey situates CloutCheck within current trends (LLMs + multimodal backbones) and provides a landscape of evaluation practices, failure modes, and suggested best practices for deployment. Use this source to frame the system design choices and to cite state-of-the-art baselines.

Data quality, bias, and ethical safeguards

Recent work on toxicity in image–text pretraining datasets underscores the risk that downstream moderation models inherit harmful biases present in training corpora. This research recommends dataset auditing, filtering, and bias-mitigation strategies for model training and fine-tuning [9]. CloutCheck will adopt these recommendations by documenting dataset curation steps, performing bias audits, and incorporating explainability mechanisms. Cite this work when detailing dataset selection, preprocessing, and ethical safeguards in the methodology.

Gaps and positioning of CloutCheck AI:

From the surveyed literature we observe several recurring gaps:

(a) most pipeline papers focus narrowly on toxicity/hate-speech and do not cover broader reputation categories such as political controversy or brand-fit;

(b) many systems address only two modalities (text + image) while fewer present full audio/video + text pipelines; (

c) there is limited work that combines multimodal detection with an interpretable reputation-scoring framework tailored for influencer/brand decisions; and

(d) real-world industrial validation and stakeholder-facing reporting are under-explored.

CloutCheck AI aims to fill these gaps by

(i) implementing a full multimodal pipeline (text, image, video, audio),

(ii) adopting robust fusion and missing-modality techniques,

(iii) broadening risk categories beyond toxicity, and

(iv) delivering explainable risk scores and professional reporting for brand decision-makers.

Project Rationale

The rapid expansion of social media has transformed the way individuals and organizations interact with the public. For brands and influencers, online presence has become a powerful tool for visibility, engagement, and revenue generation. However, this visibility also brings increased vulnerability to reputational risks, where a single post, video, or statement can severely damage credibility and long-term trust. In such a high-stakes environment, the absence of a reliable, automated, and multimodal reputation monitoring system represents a significant gap for both research and practice.

The primary purpose of **CloutCheck AI** is to address this gap by creating an intelligent platform capable of analyzing text, images, videos, and audio to provide a comprehensive reputation risk score. Unlike existing tools, which are often limited to manual review or text-only sentiment analysis, CloutCheck aims to offer an objective, data-driven, and scalable solution. This not only enhances efficiency but also minimizes human bias, ensuring more consistent and reliable assessments.

The motivation behind pursuing this project lies in its dual academic and industrial significance. From an academic standpoint, the project provides an opportunity to explore cutting-edge AI technologies in natural language processing, computer vision, and speech recognition, while integrating them into a unified decision-support framework. From an industry perspective, the solution has the potential to significantly improve brand safety, influencer marketing strategies, and risk management practices.

Through this research and development effort, we aim to deepen my understanding of multimodal AI systems, risk classification algorithms, and cloud-based software deployment. More importantly, we hope to contribute to a system that not only advances technical knowledge but also delivers practical impact by helping organizations make informed, reputation-conscious decisions in the digital age.

1.1 Aims and Objectives

The aim of this project is to design and develop an **AI-powered multimodal reputation risk analysis platform** that automatically collects and analyzes digital content from social media and online sources, evaluates reputational risks across multiple categories, and generates an objective, transparent reputation score to support informed decision-making by brands, organizations, and individuals.

To achieve the above aim, the project will focus on the following key objectives:

1. **Content Acquisition and Integration**
    - Develop automated methods for extracting user-generated content from social media platforms using APIs and custom web scrapers.
    - Ensure secure handling of data through role-based authentication and privacy-preserving mechanisms.
2. **Multimodal Content Analysis**
    - Implement natural language processing techniques to detect sentiment, hate speech, and extremism in textual data.
    - Apply computer vision models to identify high-risk elements such as nudity, drugs, and violence in images and video frames.
    - Incorporate speech-to-text transcription and offensive language detection for audio and video sources.
3. **Reputation Risk Scoring**
    - Design a transparent scoring framework that assigns severity levels (0–100) across multiple risk categories.
    - Integrate modality-specific outputs into a unified reputation index with confidence measures.

4. **Decision Support and Visualization**

    - Develop an interactive dashboard for real-time visualization of reputation metrics, trend analysis, and alerts.

    - Implement branded PDF report generation with actionable insights for decision-makers.

5. **System Deployment and Evaluation**

    - Deploy the platform on a scalable cloud-based infrastructure with support for multi-user access.

    - Evaluate system performance using accuracy benchmarks, usability testing, and comparative analysis with existing solutions.

By fulfilling these objectives, the project will deliver a functional prototype that demonstrates both the academic contribution of multimodal AI integration and the practical value of an automated reputation risk management system.

1.2 Scope of the Project

The scope of **CloutCheck AI** defines the boundaries of the work to be undertaken, outlining what the system will deliver, the features to be implemented, and the criteria for evaluating success. It also establishes the limits of the project to ensure realistic, focused, and achievable outcomes within the timeframe of a final-year project.

Project Goals

- To design and implement a multimodal reputation risk analysis platform.
- To integrate AI models for text, image, video, and audio content analysis.
- To deliver a transparent reputation scoring system that quantifies risk categories.
- To provide decision support through visualization tools and automated reporting.

Deliverables

- A fully functional prototype of the CloutCheck AI platform with a secure login system.
- Content extraction pipelines capable of integrating with selected social media platforms.
- Multimodal analysis modules for text, images, videos, and audio.
- An interactive dashboard for visualizing risk scores and generating professional reports.
- Documentation covering system architecture, methodologies, evaluation, and user guide.

Features and Functions

- Role-based user authentication.

- Automated content scraping and integration.
- Risk detection across multiple categories (hate speech, violence, nudity, extremism, drugs, etc.).
- Unified reputation index with severity scoring.
- Dashboard with visualization tools (graphs, charts, heatmaps).
- Real-time alerts and trend analysis.
- Branded PDF report generation.

Tasks

- Requirement analysis and literature review.
- Design of system architecture and data pipelines.
- Development of multimodal AI models.
- Implementation of backend and frontend components.
- Cloud deployment for scalability and access.
- Testing, evaluation, and documentation.

Deadlines

The project will follow a phased timeline:

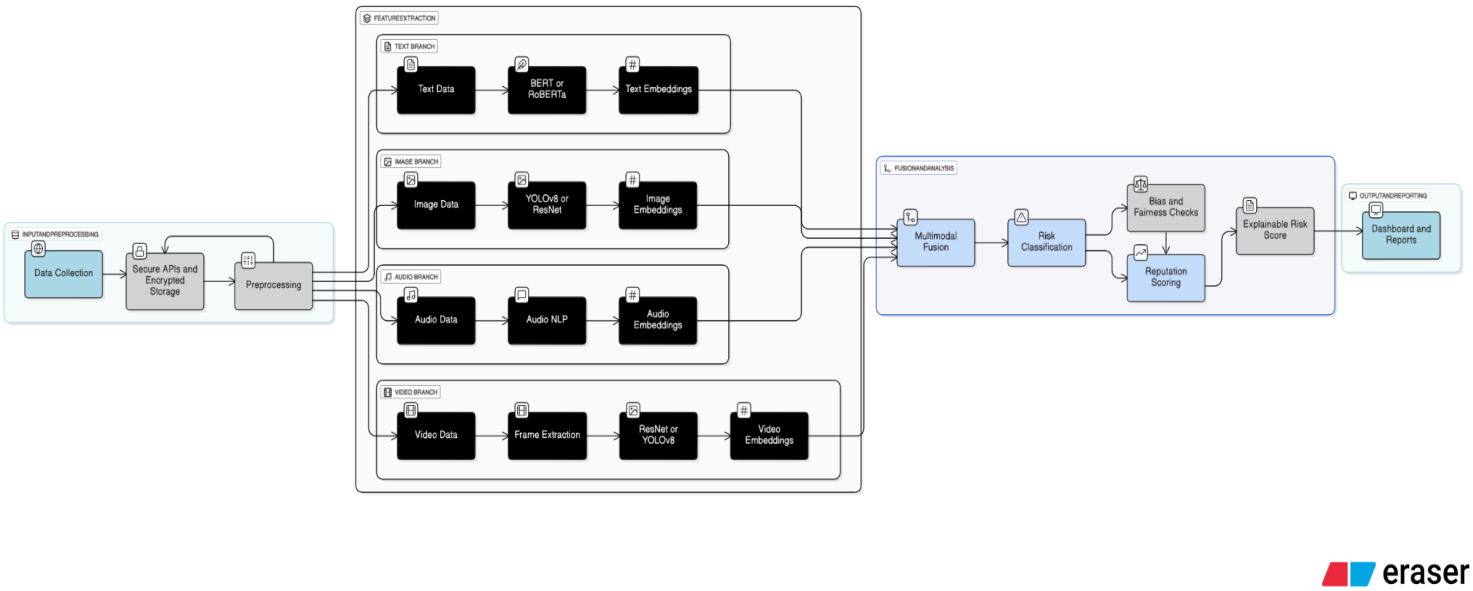
1. **Phase 1 (Months 1–3):** Requirement gathering, literature review, system design.
2. **Phase 2 (Months 3–5):** Development of scraping pipelines and multimodal analysis modules.
3. **Phase 3 (Months 5–7):** Dashboard development, integration, and report generation.
4. **Phase 4 (Months 7-9):** Testing, evaluation, and dissertation preparation.

Costs:

As an academic project, financial costs will be minimized by leveraging open-source tools, pre-trained AI models, and student-access cloud credits. However, we might need financing for external scrapers like Apify or official APIs that can save us a lot of time and dodge privacy concerns that may come with custom scraping.

The primary cost factors include external computing resources for training and deployment, which will be managed within available university or cloud free-tier limits.

Proposed Methodology and Architecture



eraser

The proposed architecture for CloutCheck integrates data collection, multimodal AI processing, and user-facing reporting into a unified system. It is structured into four main layers — data ingestion, preprocessing & feature extraction, multimodal risk classification, and user interaction & reporting — ensuring scalability, transparency, and security.

1. Data Ingestion Layer

    - Social media content is collected through official APIs and controlled scraping pipelines from platforms such as YouTube, TikTok, Twitter, and Instagram.
    - Only publicly available content is ingested to comply with data protection and platform policies.
    - The data may include text posts, captions, comments, videos, images, and audio.

2. Preprocessing & Feature Extraction Layer

    - Each modality (text, image, audio, video) undergoes cleaning and normalization.
    - Text data is processed with NLP pipelines (tokenization, stop-word removal, normalization) and encoded using BERT/RoBERTa for hate speech, toxicity, and sentiment analysis.
    - Audio and speech from videos are transcribed with Whisper, then processed as text through the NLP pipeline.
    - Image and video frames are analyzed with YOLOv8/ResNet for detection of explicit imagery, drugs, weapons, and other harmful content.
    - Each branch outputs embeddings (numerical feature vectors), representing semantic meaning for downstream fusion.

3. Multimodal Fusion & Risk Classification Layer
    - Features from different modalities are combined using fusion strategies:
        - *Early Fusion* concatenates raw features across modalities.
        - *Late Fusion* merges independent predictions via weighted voting.
        - *Attention-Guided Fusion* leverages transformer mechanisms to prioritize the most informative modality in context.
    - A unified multimodal classifier categorizes content into risk types such as hate speech, sexual/NSFW, violence, drugs/alcohol, extremism, or misinformation.
    - These risk assessments are aggregated into a Reputation Risk Score, which provides a holistic measure of an individual's online risk profile.
4. User Interaction & Reporting Layer
    - The backend services, built with FastAPI, manage orchestration of scraping, AI inference, and report generation.
    - The frontend dashboard, developed in React, provides brand managers, HR professionals, and agencies with tools for initiating scans, viewing results, and downloading PDF risk reports.
    - Reports summarize detected risks, confidence levels, and an interpretable reputation score.
5. Data & Security Layer
    - MongoDB is used to store metadata, embeddings, and scan history, while encrypted storage is employed for sensitive content and generated reports.
    - Security mechanisms include role-based access control (RBAC), audit logs, and API authentication to prevent misuse.
    - Ethical safeguards such as bias and fairness evaluation of models and explainability of risk scores are integrated into the design.

In summary, The proposed architecture ensures that CloutCheck can automatically scan multimodal digital content, detect reputational risks in real time, and present them in a user-friendly dashboard. It balances technical rigor (through fusion architectures and AI pipelines) with practical safeguards (compliance, ethics, and security), making it robust for adoption by brands, HR teams, and agencies.

Features:

1. **User Authentication & Role-Based Access**: Secure login system with role-specific permissions to ensure only authorized users can access and manage data.

2. **Profile Integration/Search**: Connects with social media APIs or allows manual entry to fetch influencer/public figure profiles for analysis.

3. **Public Data Scraping**: Uses scrapers/APIs to collect publicly available text, images, videos, and audio from different platforms.

4. **Automated Content Scraping**: Continuously gathers fresh content across platforms for real-time monitoring.

5. **Text Analysis**: Applies NLP models to detect sentiment, hate speech, and extremist content in posts and comments.

6. **Image Recognition**: Uses computer vision to identify sensitive content such as nudity, drug use, or violence.

7. **Video Frame Analysis**: Evaluates video content frame-by-frame with temporal tracking to detect harmful elements.

8. **Speech-to-Text & Toxicity Detection**: Transcribes audio/video speech and flags toxic or offensive language.

9. **Content Fusion**: Merges text, image, audio, and video insights into a unified reputation risk score.

10. **Risk Classification**: Categorizes detected risks (e.g., hate, violence, extremism) with severity levels.

11. **Real-Time Dashboard**: Provides an interactive view of risk scores, alerts, and trends for decision-makers.

12. **Agentic AI Analysis**: Explains AI findings with reasoning to improve transparency and trust.

13. **Alert System**: Sends instant notifications when high-risk content is detected.

14. **Professional Report Generation**: Creates branded, exportable PDF reports with actionable insights.

15. **Trend Analysis & Timeline**: Tracks influencer/brand risk evolution over time for long-term assessment.

16. **Cross-Platform Monitoring**: Covers multiple social media sources for comprehensive analysis.

17. **Visualization Tools**: Graphs, heatmaps, and charts to present risks in an interpretable format.

18. **Explainable AI Insights**: Provides transparency by showing why certain content was flagged.

19. **Scalable Cloud Deployment**: Ensures the platform can handle large datasets and multiple users through cloud infrastructure.

Individual Tasks

Team Member	Activity	Tentative Date
Syed Uzair (AI & Backend)	Conduct literature review on multimodal content moderation and influencer risk analysis. Identify gaps in current research.	Sep–Oct 2025
	Implement baseline AI models (NLP for text, CV for images, Speech for audio). Train and evaluate small prototypes using benchmark datasets.	Nov–Dec 2025
	Develop an improved multimodal fusion model (text + image + audio). Experiment with attention-based fusion and weak supervision. Document results.	Jan–Feb 2026
	Assist backend team with integration of AI models into APIs. Start preparing explainability features (model outputs with reasons).	Mar–Apr 2026
	Conduct evaluation study of deployed systems. Write technical chapters on AI pipeline. Support packaging, documentation, and dissertation.	May–Jun 2026
Huzaifa Khalid (Frontend & Backend)	Set up project repo, add authentication (JWT) and role-based access. Build skeleton frontend dashboard.	Sep–Oct 2025

	Design APIs for data upload and report generation. Build first prototype dashboard with sample influencer risk scores.	Nov–Dec 2025
	Extend dashboard to support multimodal visualizations (charts for sentiment, image/video analysis, audio flags). Enable PDF/export feature.	Jan–Feb 2026
	Integrate AI model APIs with frontend. Build lineage view to show how scores are computed. Optimize UI/UX.	Mar–Apr 2026
	Conduct user testing, refine dashboard, finalize product design. Package frontend/backend and contribute to dissertation.	May–Jun 2026
Saad Ahmed (AI & Backend)	Design secure data pipeline for scraping social/influencer content. Implement preprocessing flows (text cleaning, image/audio formatting).	Sep–Oct 2025
	Build scalable backend services for handling multimodal data. Add APIs for AI inference and validation reports.	Nov–Dec 2025
	Implement monitoring and anomaly detection in the backend (track failed queries, missing data). Work on Docker + CI pipeline.	Jan–Feb 2026
	Integrate all backend modules (scraping, APIs, inference, validation). Optimize performance (reduce latency, handle larger datasets).	Mar–Apr 2026
	Run system benchmarks, finalize backend documentation, support deployment and dissertation write-up.	May–Jun 2026

Gantt Chart

Tools and Technologies

The development of **CloutCheck AI** will involve a modern, scalable technology stack that ensures efficient implementation, integration of multimodal AI models, and smooth deployment. The following tools and technologies are planned for use by the project group:

1. Programming Languages

- **Python:** Primary language for backend development, AI/ML model integration, and data processing.
- **JavaScript/TypeScript:** For frontend development using React/Next.js or Angular.

2. Backend Development

- **Django (Python):** Backend web framework for handling business logic, APIs, and role-based authentication.

- **Django REST Framework (DRF):** To design and expose RESTful APIs for communication between frontend and backend.

3. Frontend Development

- **React.js / Next.js or Angular:** For building an interactive, responsive, and dynamic user interface.
- **Tailwind CSS / Material UI:** For styling, modern design, and responsive components.

4. Databases and Storage

- **PostgreSQL:** For structured relational data such as user accounts, reports, and logs.

- **MongoDB:** For unstructured or semi-structured data such as scraped metadata and JSON documents.
- **Amazon S3 / Cloud Storage:** For storing large multimedia content (images, videos, audio).

5. Data Scraping and Integration

- **Apify / Google Scrape API:** For large-scale social media content extraction.
- **Official Social Media APIs (X & Instagram etc.):** For secure and reliable profile and content integration.

6. AI / Machine Learning

- **Hugging Face Transformers:** For NLP tasks such as sentiment analysis, hate speech detection, and extremism identification.
- **CLIP, BLIP, ViT:** For image and video content classification and multimodal fusion.
- **Whisper:** For speech-to-text transcription and toxic language detection in audio content.
- **PyTorch / TensorFlow:** For training and fine-tuning custom AI models.
- **OpenAI / Groq APIs:** For generating explainable AI insights and reasoning.

7. Deployment and Infrastructure

- **Docker & Kubernetes:** For containerization, orchestration, and scaling services.
- **AWS / GCP / Azure:** For cloud hosting, compute resources, and scalability.

8. Visualization and Reporting

- **Plotly / Chart.js / D3.js:** For interactive visualization of risk scores, timelines, and heatmaps.
- **WeasyPrint / ReportLab:** For generating branded PDF reports with embedded charts and evidence.

9. Development and Collaboration Tools

- **Git & GitHub/GitLab:** For version control and collaborative development.
- **Slack / Microsoft Teams / Trello / Jira:** For project management and team communication.

- **VS Code / PyCharm:** As IDEs for coding and debugging.

References

[1] Ryu & Han (2021) - use for theoretical definition of reputation dimensions and to justify non-toxicity reputation factors (authenticity, expertise).

[2] The Dark Side of Social Media Influencers (Wiley, 2025) - use for motivational/background material on real consequences and types of reputation risk.

[3] Multimodal Post Attentive Profiling (WSDM 2020) — cite model fusion benefits (text+image) and influence-profiling methodology.

[4] ToxVidLM (ACL 2024) — cite architecture for video + speech toxicity detection, dataset construction, and evaluation metrics.

[5] MuTox (2024) — cite multilingual audio dataset and recommendations for audio toxicity detection and code-mixing handling.

[6] Rethinking Multimodal Content Moderation (AM3, 2023) — cite asymmetric modality analysis and modality-weighting strategies for fusion.

[7] Multimodal Guidance Network (2023) — cite missing-modality inference methods and guidance networks for robust predictions.

[8] Recent Advances in Hate Speech Moderation (2024) — cited as a survey of multimodal + LLM trends and evaluation best practices.

[9] Understanding & Mitigating Toxicity in Image-Text Pretraining Datasets (2025) — cite for dataset auditing, bias mitigation, and ethical safeguards.