CYBERSHIELD HACKATHON

TOPIC:-Tools that Protect Women Online (Privacy, Stalking Defense, Harassment Detection)

Team Name:- **GODFATHER**

Team Members:- KAVYA GUPTA,RAVI KANT MISHRA,AKSHAY KUMAR MISHRA,ANUBHAVI JAISWAL

IDEA TITLE

IDEA/SOLUTION :

Development of a **Digital Mental Health and Psychological Support System** for Students in Higher Education

- **AI-Powered Moderation Engine** – Detects toxicity in text (comments, emails, posts) using **Bytez.js model with retry logic ,Queue based system** and Redis caching for efficiency.
- **Multi-Modal Content Filtering** – Image moderation via **Sightengine API** to flag nudity, gore, self-harm, weapons, and offensive visuals.
- **Browser Extension Integration** – Real-time scanning on Gmail, Instagram, Twitter/X, and Facebook through content scripts and background workers.
- **Cross-Platform Moderation API** – Express.js backend with dedicated routes (/moderation) for handling text, image, and comment analysis.
- **Caching & Performance Optimization** – **REDIS** caching ensures repeated toxic content checks are instant, reducing API calls and latency.

PROBLEM RESOLUTION :

- **Automated Abuse Detection & Protection** – **Detects toxic, harmful, or offensive content** in real-time across platforms (Gmail, Instagram, Facebook, X), reducing exposure to harmful material and protecting users seamlessly.
- **Scalable & Efficient Moderation** – **Eliminates manual review bottlenecks** with automation, while caching and retry mechanisms ensure cost-effective, large-scale deployment without delays.

UVP (Unique Value Proposition) :

- **All-in-One Safety Layer** – First **lightweight extension** combining text + **image moderation** in real-time across multiple social platforms.
- **High Accuracy & Speed** – OpenSource AI models (Bytez.js + Sightengine) with Redis caching deliver fast and reliable toxicity detection.
- **Plug-and-Play Integration** – Works as a Chrome extension, requiring no platform-side modification.
- **Privacy-Conscious Architecture** – Data processed securely with caching limited to temporary 5-minute windows, minimizing storage of user content.
- **Scalable & Adaptable** – Can be extended to additional platforms and moderation categories (e.g., cyberbullying, scam detection).

TECHNICAL APPROACH

BACKEND & API DEVELOPMENT

- Node.js & Express.js – REST API framework for scalable content moderation services
- Bytez API Integration – AI-powered text toxicity detection with retry and caching mechanisms
- Sightengine API – Image moderation with nudity, violence, gore, and offensive content detection
- Redis Caching – High-speed cache to optimize moderation results and reduce repeated model calls

BROWSER EXTENSION & WEB INTEGRATION

- Chrome Extension (Manifest v3) – Real-time moderation on Gmail, Instagram, Twitter (X), and Facebook
- Content Scripts – Automatically scan comments, messages, and emails for harmful or toxic content
- Popup UI – Lightweight interface for quick insights on flagged content

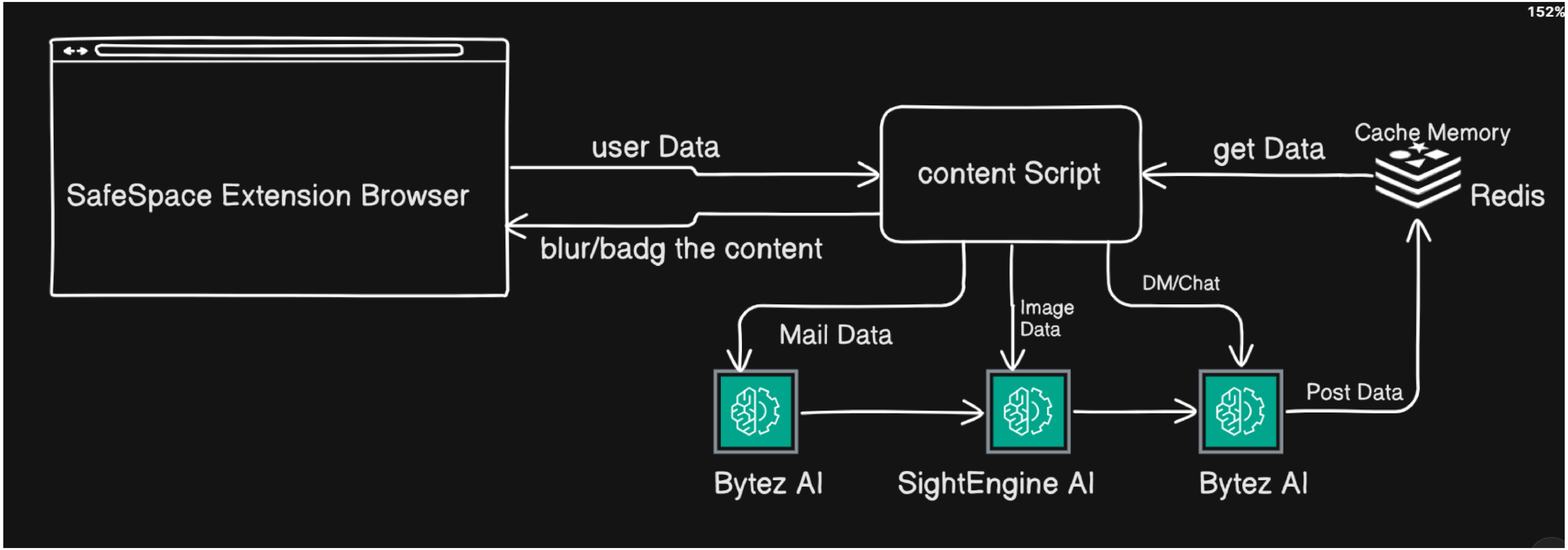
DATABASE & PERFORMANCE

- Redis – In-memory data store for caching moderation results and reducing API overhead

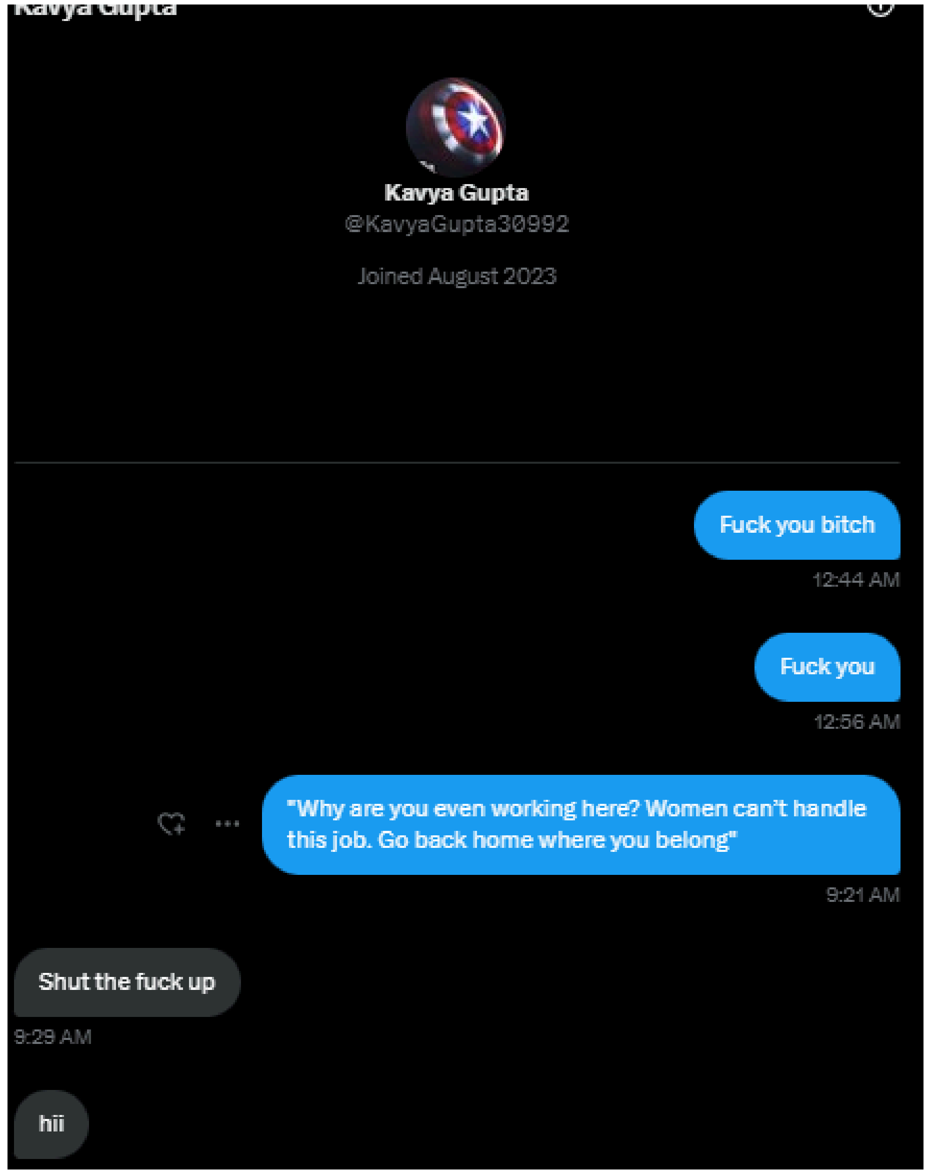
INTEGRATION & SECURITY

- SHA-256 Hashing – Unique cache keys for secure and collision-free content lookups
- CORS Enabled APIs – Safe cross-origin requests from the extension to backend services

Architecture Diagram



Chats/Dm moderation



Kavya Gupta

Kavya Gupta
@KavyaGupta30992
Joined August 2023

Fuck you bitch
12:44 AM

Fuck you
12:56 AM

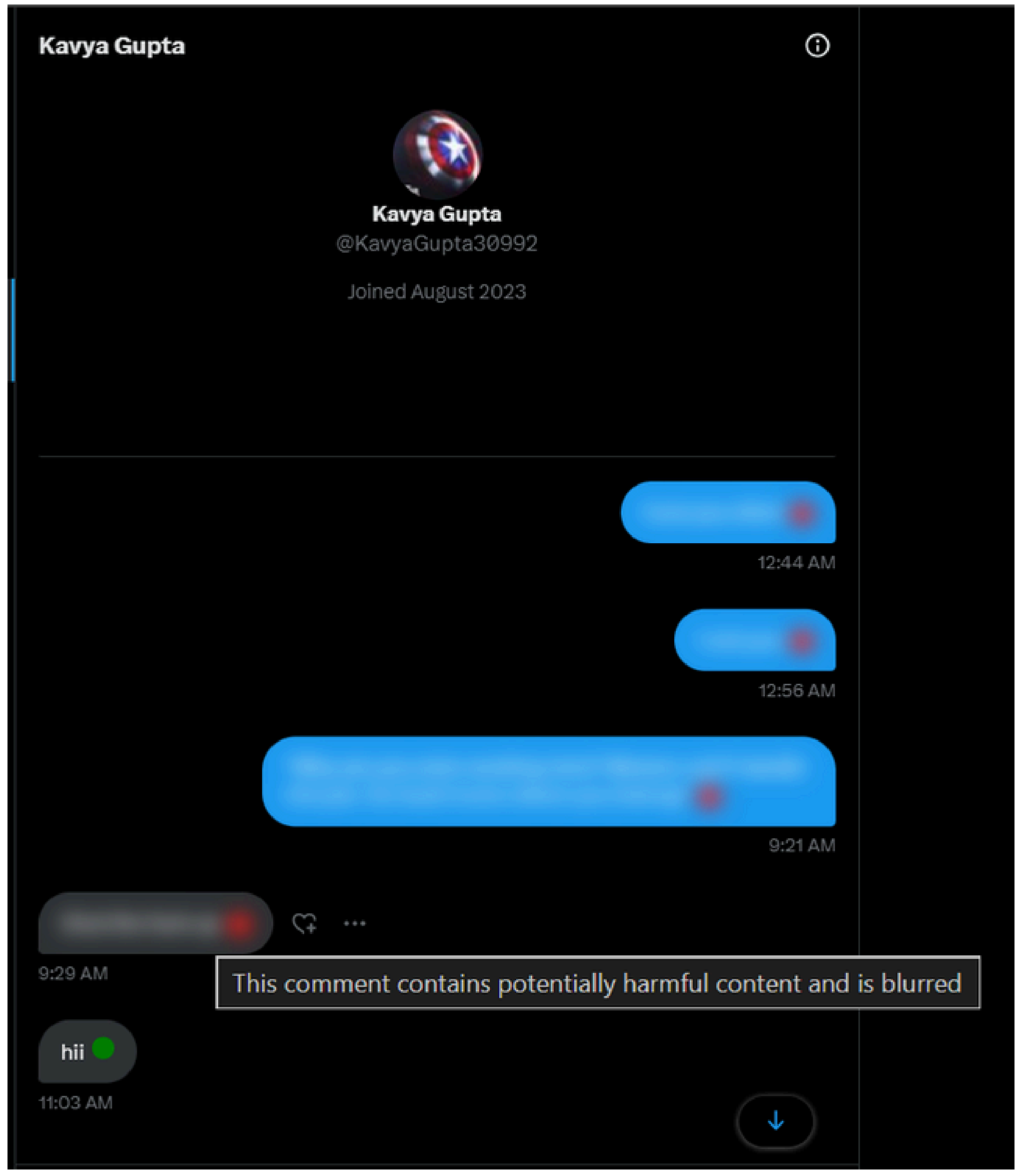
"Why are you even working here? Women can't handle this job. Go back home where you belong"
9:21 AM

Shut the fuck up
9:29 AM

hii

Harrasment

Moderation



Kavya Gupta

Kavya Gupta
@KavyaGupta30992
Joined August 2023

12:44 AM

12:56 AM

9:21 AM

9:29 AM

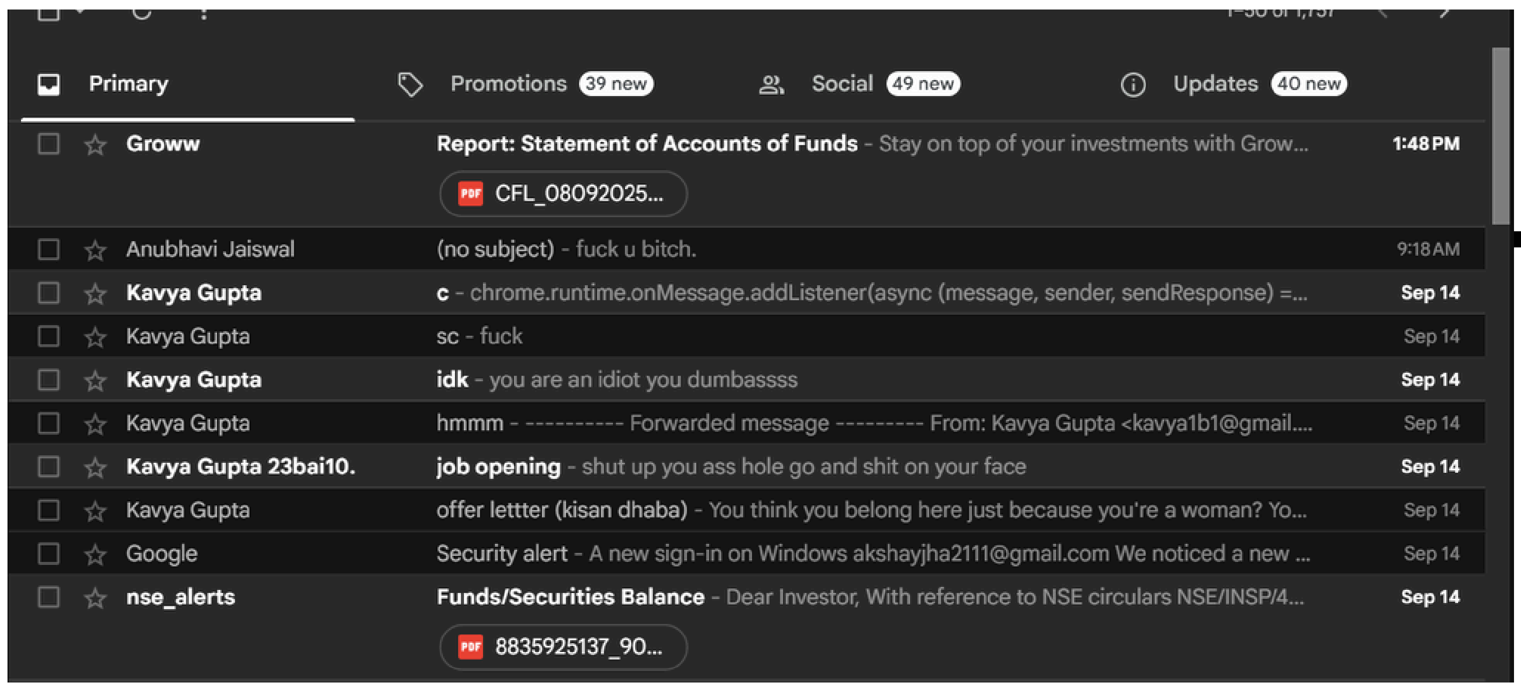
This comment contains potentially harmful content and is blurred

hii
11:03 AM

Before

After

Email Moderation



1–50 of 1,757

Primary | Promotions 39 new | Social 49 new | Updates 40 new

Groww — Report: Statement of Accounts of Funds - Stay on top of your investments with Grow... — 1:48 PM
CFL_08092025...

Anubhavi Jaiswal — (no subject) - fuck u bitch. — 9:18 AM

Kavya Gupta — c - chrome.runtime.onMessage.addListener(async (message, sender, sendResponse) =... — Sep 14

Kavya Gupta — sc - fuck — Sep 14

Kavya Gupta — idk - you are an idiot you dumbassss — Sep 14

Kavya Gupta — hmmm - ---------- Forwarded message --------- From: Kavya Gupta <kavya1b1@gmail.... — Sep 14

Kavya Gupta 23bai10. — job opening - shut up you ass hole go and shit on your face — Sep 14

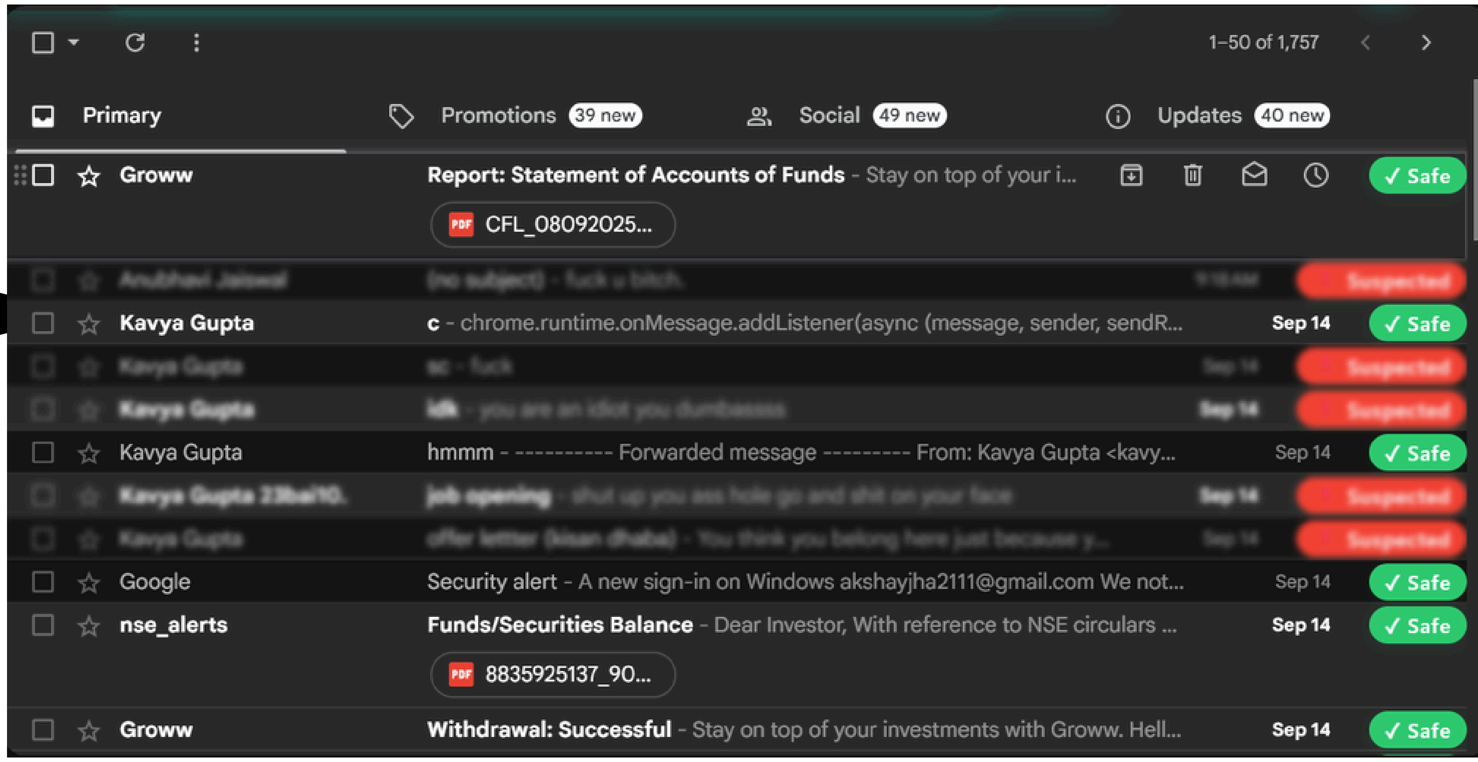
Kavya Gupta — offer lettter (kisan dhaba) - You think you belong here just because you're a woman? Yo... — Sep 14

Google — Security alert - A new sign-in on Windows akshayjha2111@gmail.com We noticed a new ... — Sep 14

nse_alerts — Funds/Securities Balance - Dear Investor, With reference to NSE circulars NSE/INSP/4... — Sep 14
8835925137_90...

Harrasment

Moderation



1–50 of 1,757

Primary | Promotions 39 new | Social 49 new | Updates 40 new

Groww — Report: Statement of Accounts of Funds - Stay on top of your i... — ✓ Safe
CFL_08092025...

Anubhavi Jaiswal — (no subject) - fuck u bitch. — 9:18 AM — Suspected

Kavya Gupta — c - chrome.runtime.onMessage.addListener(async (message, sender, sendR... — Sep 14 — ✓ Safe

Kavya Gupta — sc - fuck — Sep 14 — Suspected

Kavya Gupta — idk - you are an idiot you dumbassss — Sep 14 — Suspected

Kavya Gupta — hmmm - ---------- Forwarded message --------- From: Kavya Gupta <kavy... — Sep 14 — ✓ Safe

Kavya Gupta 23bai10. — job opening - shut up you ass hole go and shit on your face — Sep 14 — Suspected

Kavya Gupta — offer lettter (kisan dhaba) - You think you belong here just because y... — Sep 14 — Suspected

Google — Security alert - A new sign-in on Windows akshayjha2111@gmail.com We not... — Sep 14 — ✓ Safe

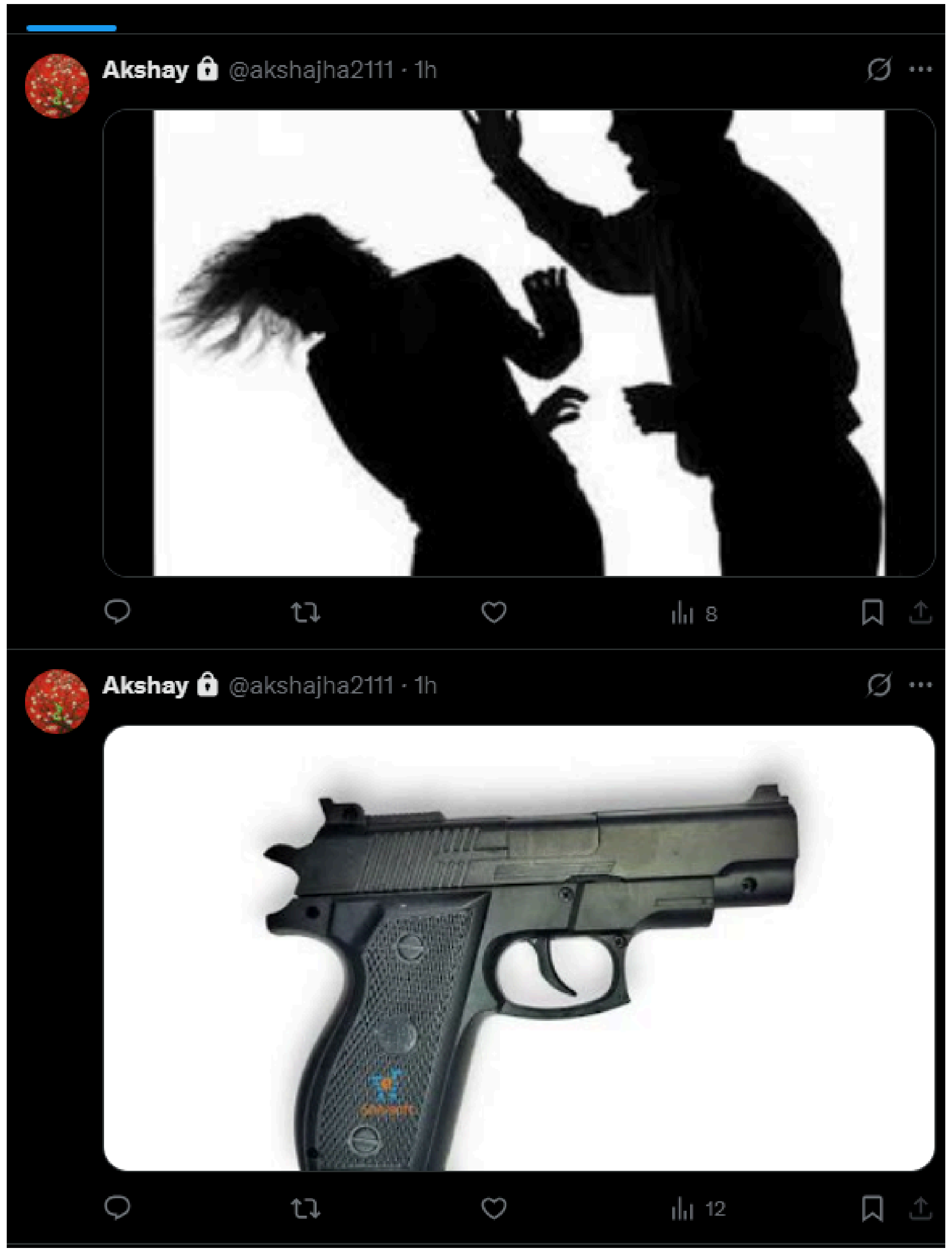
nse_alerts — Funds/Securities Balance - Dear Investor, With reference to NSE circulars ... — Sep 14 — ✓ Safe
8835925137_90...

Groww — Withdrawal: Successful - Stay on top of your investments with Groww. Hell... — Sep 14 — ✓ Safe

Before

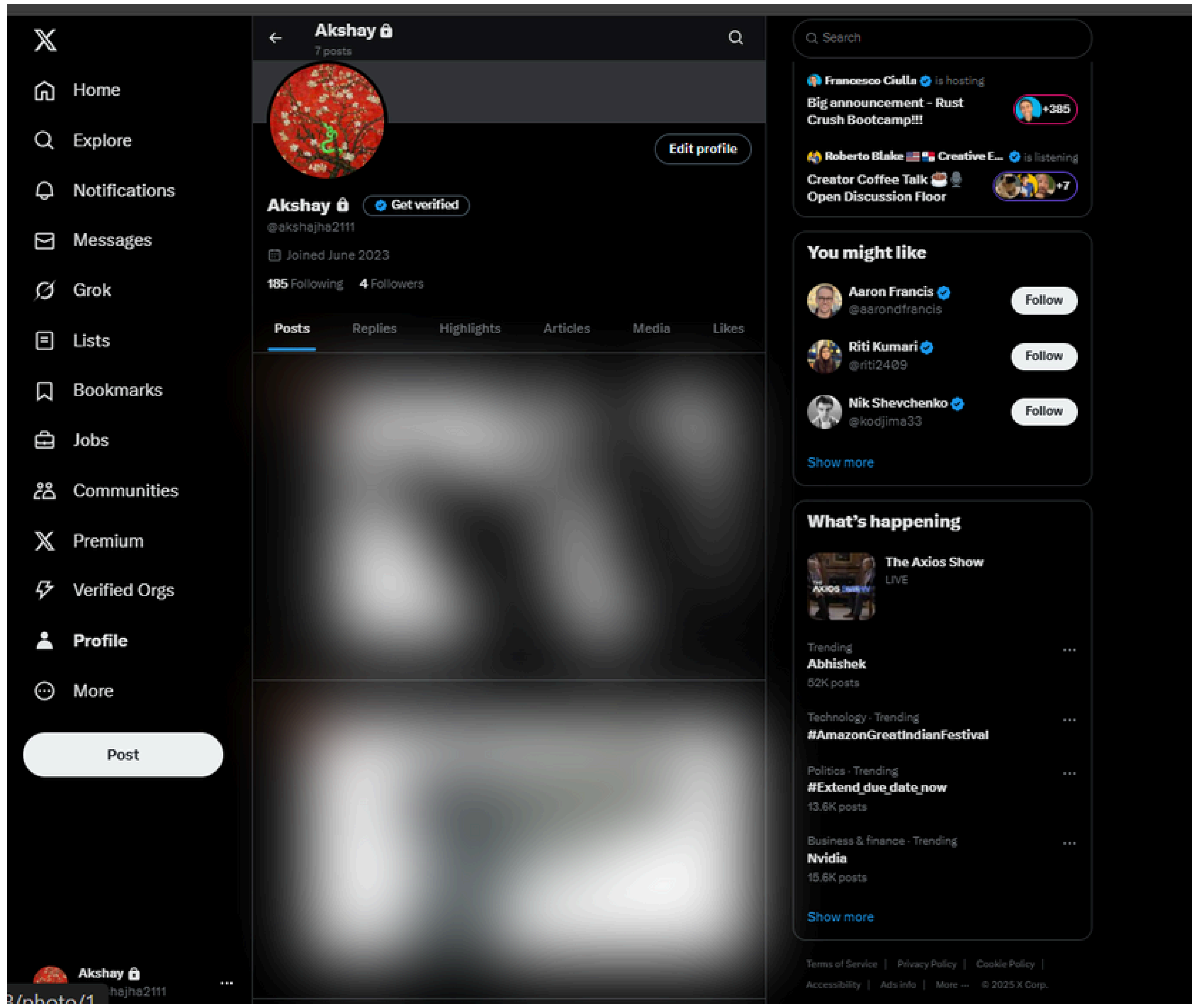
After

Post Moderation



Harrasment

Moderation



Before

After

FEASIBILITY AND VIABILITY

Feasibility

1. **High Demand**: Rising concerns over online toxicity, cyberbullying, and harmful content across social media and emails.
2. **Existing Tools:** Leverages proven APIs — Bytez (text toxicity detection) and Sightengine (image moderation).
3. **Implementation:** Backend built on Node.js + Express.js with Redis caching; Chrome extension integrates moderation into Gmail, Instagram, Facebook, and Twitter/X.
4. **Cost Structure:** Cloud-hosted services and caching reduce API overhead; scalable architecture for future integration.

Viability

1. **Impact Potential:** Helps institutions, workplaces, and students maintain safe digital spaces by detecting toxic or harmful content in real time.
2. **Institutional** Need: Deployable at organizational, campus, or enterprise levels; aligns with policies on digital safety and online well-being.
3. **Growth Trajectory:** Expands from browser extension to mobile/web apps; can integrate analytics dashboards for administrators.
4. **Market Gap:** Fills the need for lightweight, AI-driven, real-time moderation tools without relying on costly enterprise-only solutions.

Challenges

1. Technical Challenges: Handling API rate limits, concurrency restrictions (Bytez), and ensuring low-latency moderation.
2. Data Privacy: Secure handling of user text/images; compliance with data protection standards.
3. Coverage: Adapting moderation for multilingual and context-sensitive toxicity detection.
4. User Adoption: Balancing moderation accuracy with user trust; avoiding false positives that may disrupt normal use.

Solutions

1. Technology Adoption: Lightweight Chrome extension for instant integration into existing platforms.
2. Performance Optimization: Redis caching and SHA-256 hashing ensure fast, secure, and efficient lookups.
3. User Engagement: Clear feedback (ratings/messages) for flagged content to guide user behavior.
4. Future Enhancements: Expansion to mobile platforms, role-based moderation dashboards, and institution-level analytics.

IMPACT AND BENEFITS

IMPACT ON STAKE HOLDERS

User Safety & Experience

- Safer online communication on Gmail, Instagram, Facebook, and Twitter/X.
- Real-time flagging of toxic comments, harmful text, and unsafe images.

Market & Social Benefits

- Helps reduce cyberbullying, harassment, and exposure to harmful content.
- Promotes healthier digital spaces across educational and professional environments.

Institutional Performance

- Assists schools, colleges, and organizations in monitoring online interactions.
- Provides reliable moderation support without costly enterprise-only tools.

Developer & Admin Efficiency

- Saves time by caching results with Redis, reducing API calls and costs.
- Provides scalable, AI-powered moderation APIs for easy integration.

Benefits of Our Solution

- **Real-Time Moderation:** Instantly detects and flags toxic text or unsafe images in user interactions.
- **Efficient Caching & Performance:** Redis + SHA-256 hashing ensures fast responses and reduced server load.
- **Seamless Integration:** Works directly via Chrome Extension on major platforms (Gmail, Instagram, Facebook, Twitter/X).
- **AI-Powered Detection:** Uses Bytez API for text toxicity and Sightengine API for harmful image detection.
- **Secure & Scalable:** Cloud-ready design with reliable API endpoints, extendable to mobile/web apps in future.

RESEARCH AND REFERENCES

TECHNICAL DOCUMENTATION

- BYTEZ API – TOXICITY CLASSIFICATION MODELS AND CONCURRENCY HANDLING IN MODERATION PIPELINES.
- REDIS CACHING GUIDE – SHA-256 CACHE KEYS, TTL, AND JSON RESULT CACHING FOR LOW-LATENCY APIS**.**

MARKET RESEARCH SOURCES

- FORTUNE BUSINESS INSIGHTS (2024) – GROWTH IN CONTENT MODERATION DUE TO RISING DIGITAL ENGAGEMENT.
- MARKETWATCH (2024) – CHALLENGES: FALSE POSITIVES, CULTURAL CONTEXT, AND MAINTAINING USER TRUST.

ACADEMIC & TECHNICAL REFERENCES

- IEEE TRANSACTIONS ON AFFECTIVE COMPUTING – NLP AND DEEP LEARNING FOR DETECTING ONLINE HARASSMENT.
- ACM DIGITAL LIBRARY – IMAGE-BASED MODERATION WITH CNNS AND MULTIMODAL AI APPROACHES.