< PU CODE
HACKATHON 3.0 >

Parul® University | NAAC GRADE A++

<PU CODE HACKATHON 3.0 2026>

Fields	Information
Problem Statement Title	AI Powered Online Harassment Detector
Team Name	Cipher Trace
Team Leader Name	Jyanesh Naidu
Institute Name	Parul Institute of Engineering and Technology
Track Name	Cyber Security
Team Member	Manish Kashyap, Susritha Swamyvari ,Koushik Yadav

Proposed Solution
AI-Powered Online Harassment Detection System

❌ **Problem**	▪ Rapid rise in online harassment and cyberbullying ▪ Manual moderation is slow and not scalable ▪ Harmful impact on users’ mental health ▪ Lack of real-time detection systems
🎯 **Objective**	▪ Detect online harassment using AI in real time ▪ Classify abuse types and severity ▪ Prevent repeated harassment incidents ▪ Ensure safer digital platforms
⚙️ **Approach**	▪ Analyze user messages using NLP techniques ▪ Apply AI models for harassment classification ▪ Generate risk score for severity detection ▪ Take automated actions (warn, block, report)

USE CASES

- 👥 Social Media Platforms
- 🎓 Online Education Platforms
- 🎮 Gaming & Live Streaming Platforms
- 🏢 Corporate Communication Tools
- 🧠 Mental Health & Child Safety

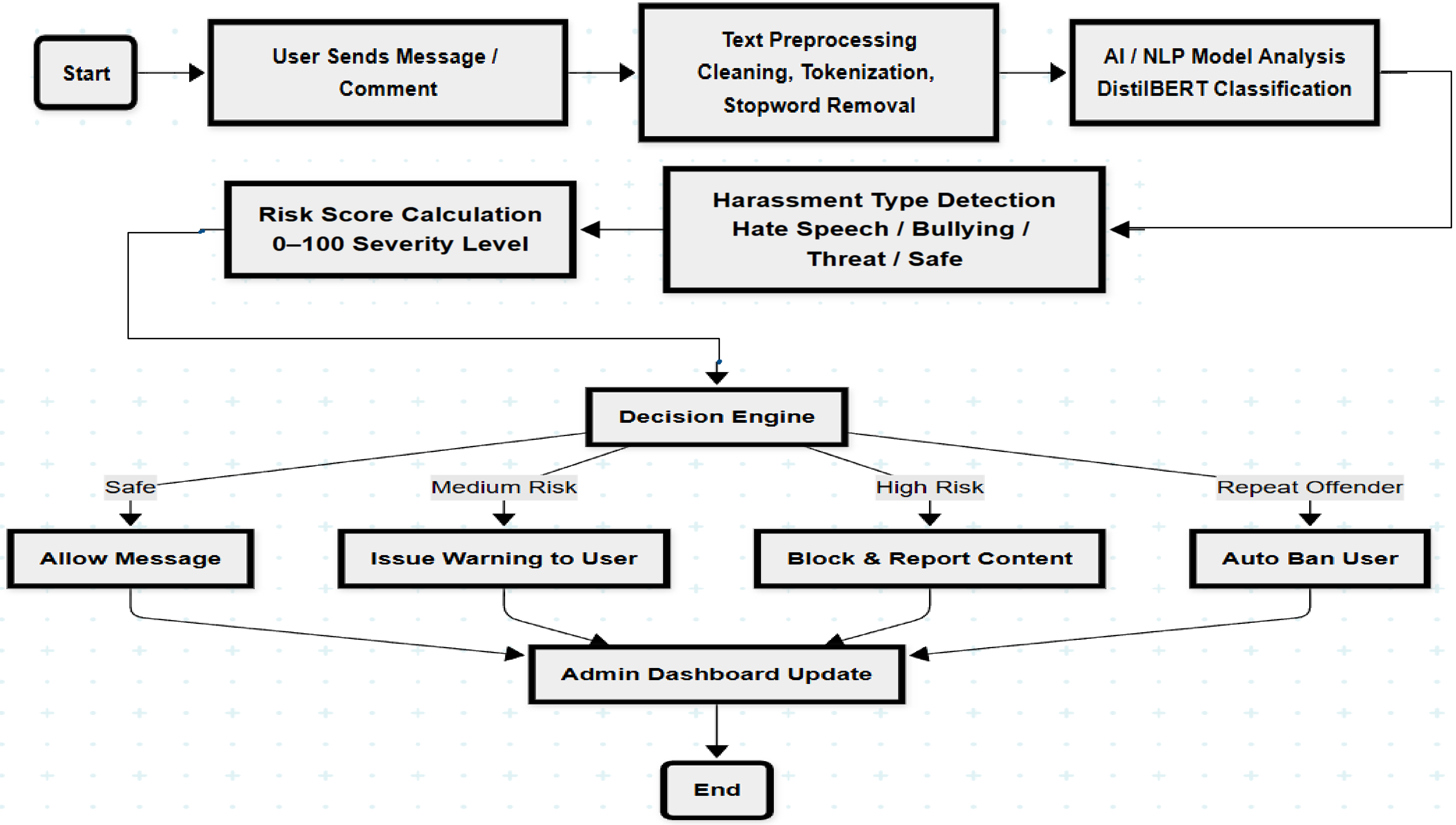
TECHNOLOGY STACK

🤖 **AI / Machine Learning**	▪ Python ▪ DistilBERT (NLP Model) ▪ HuggingFace Transformers ▪ Scikit-learn
⚙️ **Backend Services**	▪ FastAPI ▪ REST APIs ▪ JWT Authentication
🖥️ **Frontend & Dashboard**	▪ React.js / Streamlit ▪ HTML, CSS, JavaScript
🗄️ **Database & Storage**	▪ MongoDB ▪ JSON-based logs
🔐 **Security & Monitoring**	▪ Role-based access control ▪ Secure API communication
☁️ **Deployment**	▪ Render / Railway ▪ Docker

< PU CODE HACKATHON 3.0 >

PROCESS FLOWCHART

Parul® University | NAAC A++ GRADE


Start
User Sends Message / Comment
Text Preprocessing Cleaning, Tokenization, Stopword Removal
AI / NLP Model Analysis DistilBERT Classification
Harassment Type Detection Hate Speech / Bullying / Threat / Safe
Risk Score Calculation 0–100 Severity Level
Decision Engine
Safe
Medium Risk
High Risk
Repeat Offender
Allow Message
Issue Warning to User
Block & Report Content
Auto Ban User
Admin Dashboard Update
End

DEPENDENCIES

❑ Limited availability of high-quality labeled harassment datasets

❑ Difficulty handling multilingual content, slang, and evolving abusive language

❑ Challenges in detecting sarcasm and context-based harassment

❑ High computational requirements for real-time moderation

❑ Risk of false positives affecting genuine users

❑ Need for continuous model updates to track new abuse patterns

Mitigation Approach

- Use pre-trained transformer models
- Continuous learning using feedback loop
- Multilingual model support
- Human-in-the-loop moderation for edge cases