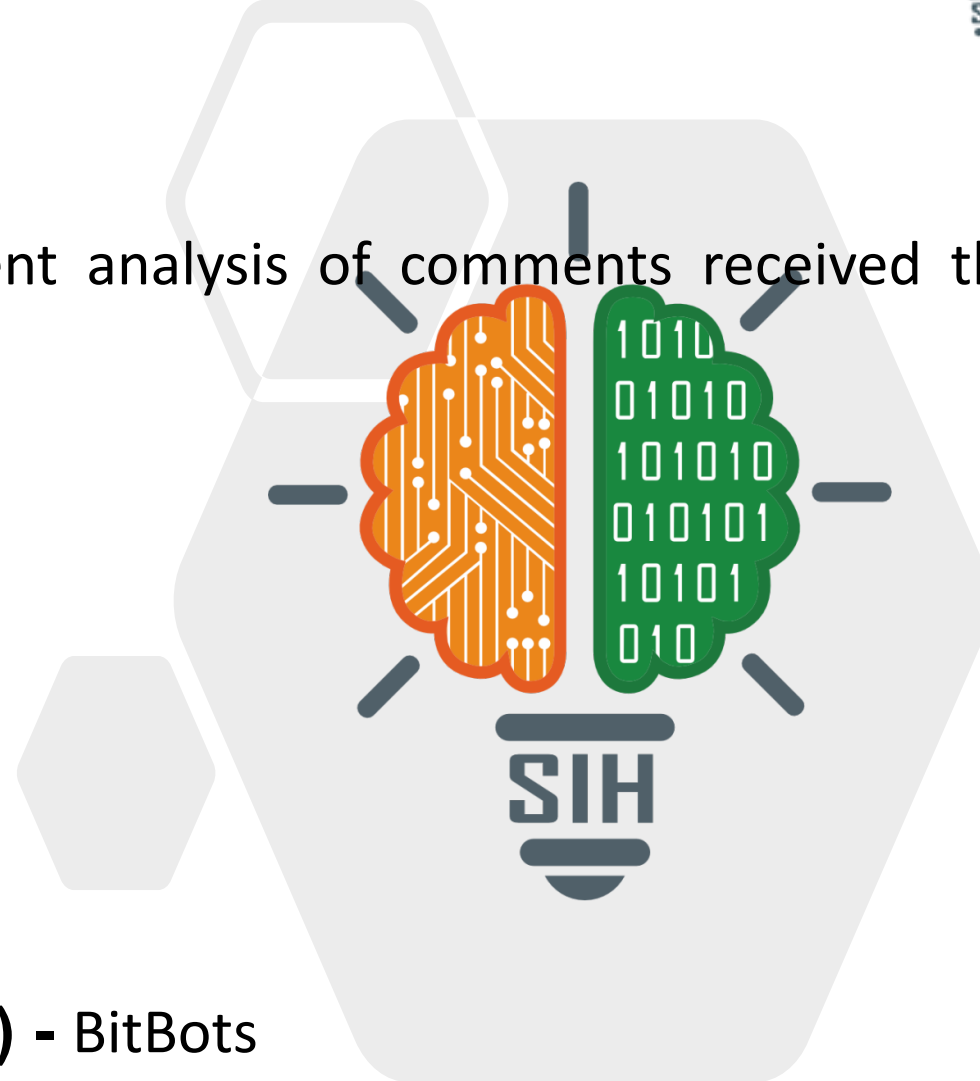


SMART INDIA HACKATHON 2025



- **Problem Statement ID** – 25035
- **Problem Statement Title-** Sentiment analysis of comments received through E-consultation module
- **Theme-** Miscellaneous
- **PS Category-** Software
- **Team ID-**
- **Team Name (Registered on portal) -** BitBots



IDEA TITLE

❖ Proposed Solution

- **API Base**
 - Collect comments/reviews from MCA website and store them in a vector DB using multilingual embeddings + re-ranker and multilingual output using NLP.
 - Comment prioritization using Re-ranker and impact score.
- **Structured Storage & Analysis**
 - Save with metadata (comment text, date, amendment name, sentiment, key words) → enabling filtering by date/amendment/sentiment.
- **Insights & Innovation**
 - Provide sentiment analysis, summary generation, and word cloud.
 - Use RAG for context-aware responses and advanced sentiment analysis across different filters.
 - Implementation of XAI – Explainable ai to explain reason of +ve or –ve responses

Technologies to be Used

Backend

- **Python** – NLTK, Huggingface, Unsloth, Langchain and FastAPI for RestApi, ingestion & analysis.
- **Fine-tuned LLM** - for sentiment analysis & summarization.
- **Sentence Transformers + Reranker (bge-reranker-v2-m3)** - multilingual embeddings & contextual ranking.
- **Vector DB (Chroma DB)** - store comments with metadata (date, amendment, sentiment, keywords).
- **Postgres DB** - store raw data from MCA website.
- **RAG Pipeline** - ensures long-term contextually correct responses.

Frontend

- **React.js** - interactive dashboards & filtering.
- **Chart.js / Recharts** - sentiment distribution & trend graphs.
- **Wordcloud.js** - keyword density visualization.

Other Tools

- **Docker, GitHub Actions** - CI/CD.
- **AWS & Cloudflare** – Hosting & CDN for fast response and caching for easy scalability.

Methodology and process for implementation

Data Collection

- Comments collection from MCA website.

Processing Pipeline

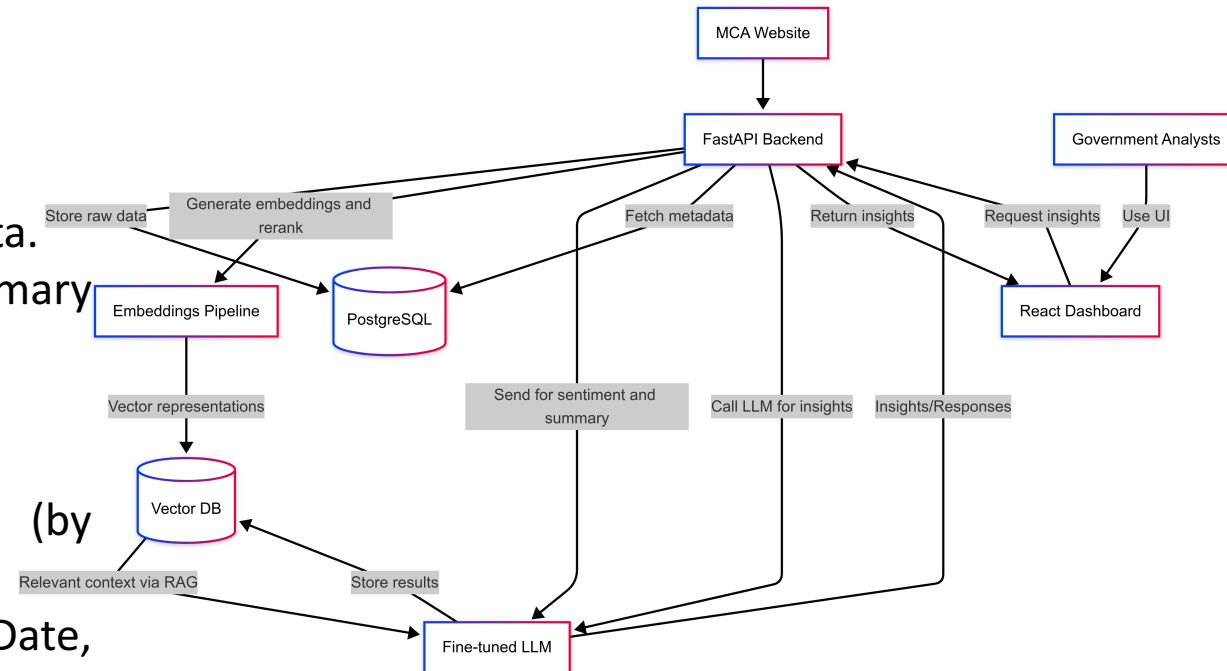
- Embeddings + Reranker → store in Vector DB with metadata.
- Fine-tuned LLM → sentiment classification + summary generation.

Knowledge Store

- Vector DB enables semantic search & filtering (by amendment/date/sentiment).
- Postgres DB for storing raw data from MCA website (Date, amendment, comments) for later use on Frontend.

Insights & Visualization

- React dashboard → word clouds, sentiment graphs, summaries, filtering options.



Feasibility Analysis

- Technically feasible – Multilingual transformer models, sentiment analyzers, and vector DB (ChromaDB) are mature and widely used.
- Operationally feasible – Automates manual review of thousands of comments, reducing analyst workload.
- Economically feasible – Open-source models and scalable cloud infra keep costs manageable.

Potential Challenges & Risks

- Language complexity – Mixed languages, abbreviations, and legal jargon may reduce accuracy.
- Performance issues – Heavy comment loads during major amendments may cause slow processing.
- Bias & misinterpretation – Sentiment models may misread sarcasm or neutral legal feedback.

Strategies to Overcome Challenges

- Fine-tune models on domain-specific (legal/consultation) datasets and add human-in-the-loop validation.
- Docker based deployment with horizontal scaling with load balancers and Optimize for scale using batch processing, caching, and lightweight quantized models.
- Improve reliability with rule-based keyword checks, confidence scores, and interpretability dashboards.

Potential Impact on the Target Audience

- Policy Makers & Analysts – Faster, more reliable insights from thousands of comments without manual overload.
- Stakeholders (Citizens, Industry, Professionals) – Assurance that their feedback is systematically considered and not overlooked.
- Government Efficiency – Transparent, data-driven decision-making leading to better governance and trust.

Benefits of the Solution

- Social
 - Promotes inclusivity by analyzing multilingual feedback fairly.
 - Increases transparency and citizen trust in government processes.
- Economic
 - Reduces manpower and time costs for analyzing bulk feedback.
 - Enables quicker policy decisions, leading to faster implementation of reforms.
- Environmental
 - Digital AI-driven analysis reduces reliance on physical paperwork and manual processing.
 - Efficient systems minimize resource usage for repetitive administrative tasks.

- MCA Website: <https://www.mca.gov.in/content/mca/global/en/home.html>
- FlowChat Diagram: <https://mermaid.live>
- OpenSource Models: <https://huggingface.co>
- AI: ChatGPT, Gemini & Perplexity
- NLTK, HuggingFace, Unsloth & Langchain documentations