



PROBLEM STATEMENT ID : PS 2

PROBLEM STATEMENT : WEB FORUM THREAT INTELLIGENCE

SCRAPER AND ANALYZER

TEAM NAME : LUPIN MEMBERS

TEAM ID : HK - 187

TEAM MEMBERS :

DEEPAK

PANKAJ

SUNNY

PALAK





PROBLEM & SOLUTION

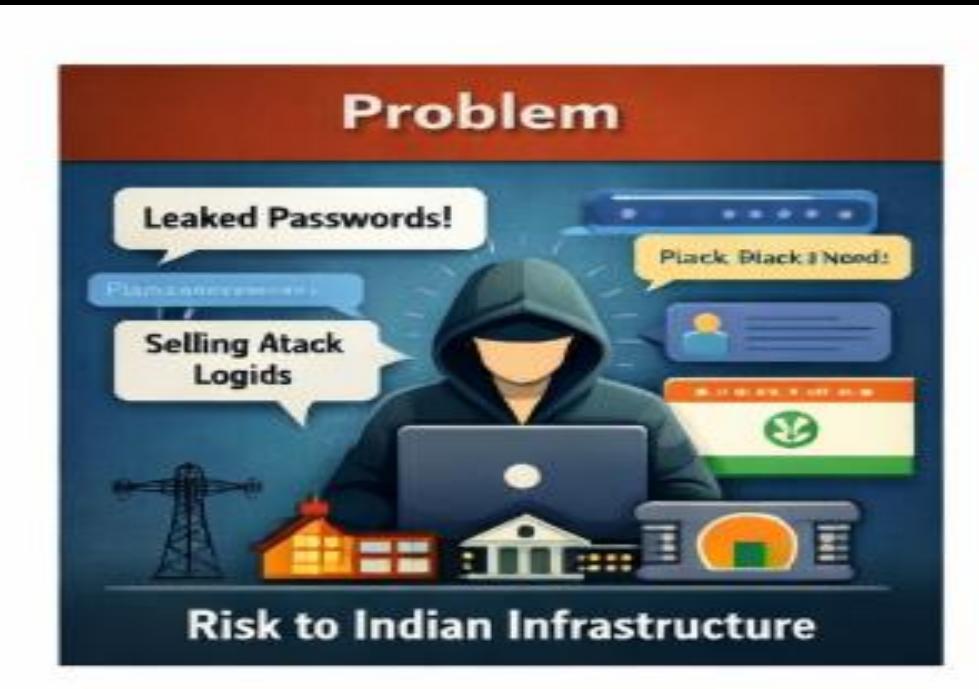
Problem :

Many cyber-attacks and data leaks are discussed or planned on online forums before they actually happen. Hackers sometimes share stolen usernames, passwords, or talk about attacking banks, power grids, or government systems. Currently, organizations cannot manually monitor thousands of forum posts every day, which leads to delayed detection of threats.

Solution :

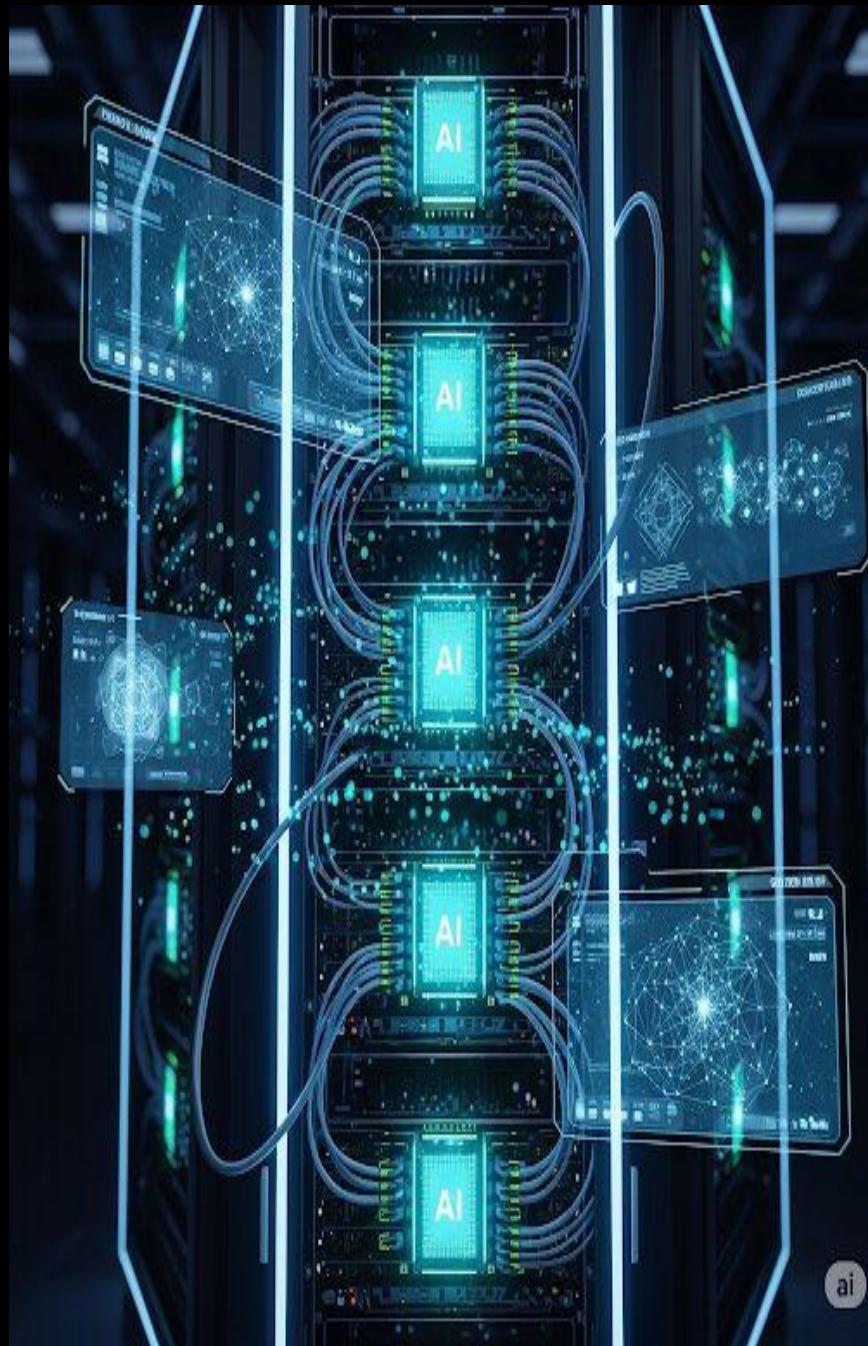
Our project builds an automated threat intelligence system that continuously scans selected online forums, detects leaked credentials, and identifies discussions related to cyber-attacks on Indian infrastructure. The system automatically analyzes posts and generates alerts so security teams can act early instead of reacting after damage is done.

Our project builds an automated threat intelligence system that continuously scans selected online forums, detects leaked credentials, and identifies discussions related to cyber-attacks on Indian infrastructure. The system automatically analyzes posts and generates alerts so security teams can act early instead of reacting after damage is done.





FLOW OF SOLUTION



- **Data Collection** - The system automatically scrapes selected web forums at scheduled intervals.
- **Data Cleaning** - Removes unnecessary symbols, duplicates, and irrelevant text.
- **Content Analysis** - Uses NLP and keyword detection to understand the meaning of posts.
- **Credential Detection** - Finds patterns like email-password pairs, API keys, or config files.
- **Threat Classification** - Assigns severity levels such as Low, Medium, High, or Critical.
- **Log Storage** - Saves historical data for future trend analysis.



TECH STACK & APPROACH

Tech Stack :

- **Programming Language:** Python
- **NLP Libraries:** NLTK / SpaCy / Transformers
- **Database:** MySQL / MongoDB
- **Dashboard:** React / Flask / Django
- **Scheduling:** Cron Jobs / Task Scheduler



Approach :

We follow a modular approach where each module (scraper, analyzer, database, dashboard) works independently but connects to form a complete automated monitoring pipeline.

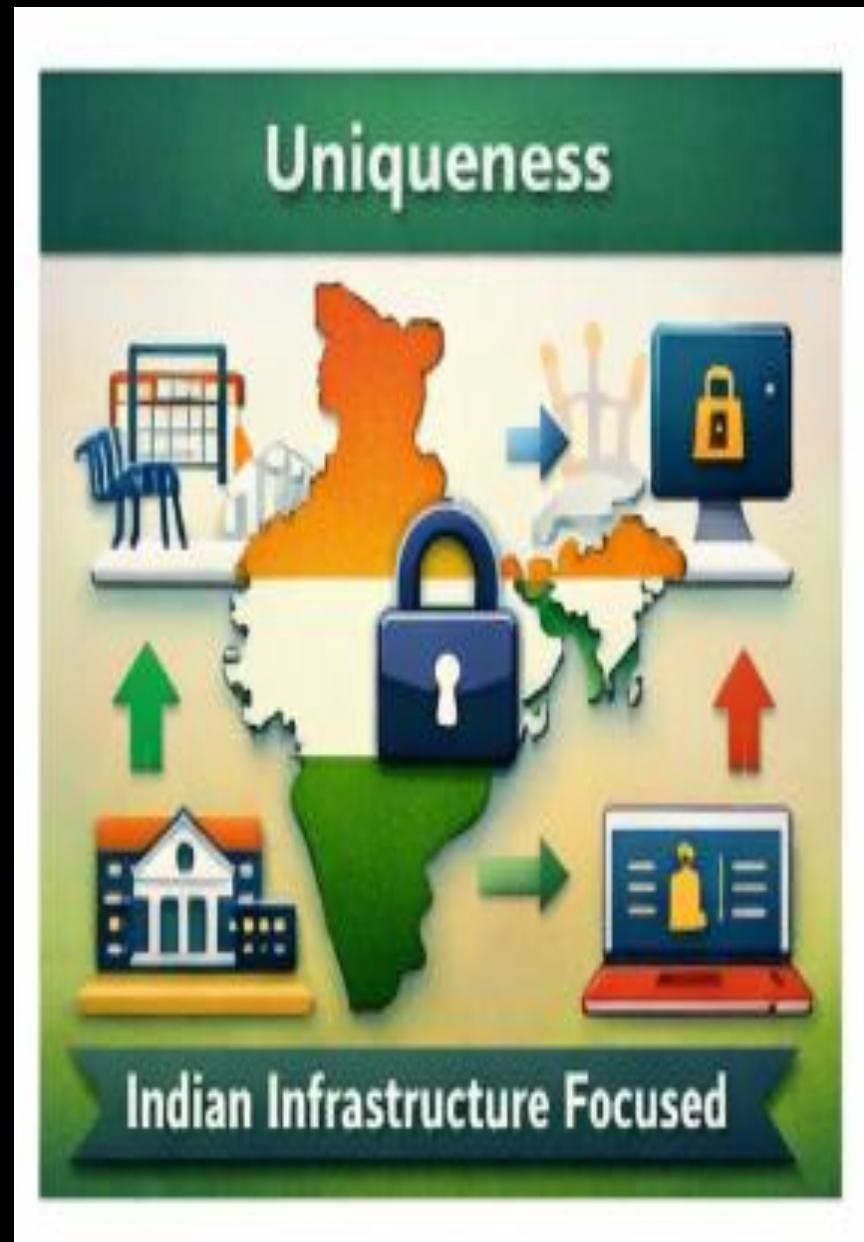




UNIQUENESS & INNOVATION FACTOR

Uniqueness and Innovation Factor:

- **Integrated Multi-Layer Intelligence System** - Combines web scraping, NLP, pattern detection, and threat scoring in one platform. This provides deeper analysis compared to basic keyword-based tools.
- **India-Focused Critical Infrastructure Monitoring** - Specifically targets sectors like banking, power, telecom, and government services. This makes the system more relevant and practical for national cybersecurity needs.
- **Context-Aware Threat Understanding** - Analyzes the meaning and intent behind forum discussions instead of only matching words. This reduces false positives and improves detection accuracy.
- **Automated Threat Prioritization and Alerting** - Uses a scoring model to rank threats based on severity, credibility, and impact. Security teams can quickly focus on the most critical risks.





FEASIBILITY & CHALLENGES

Feasibility :

- Uses widely available open-source tools and libraries.
- Can be built with moderate computing resources.
- Modular design allows easy upgrades and scaling.



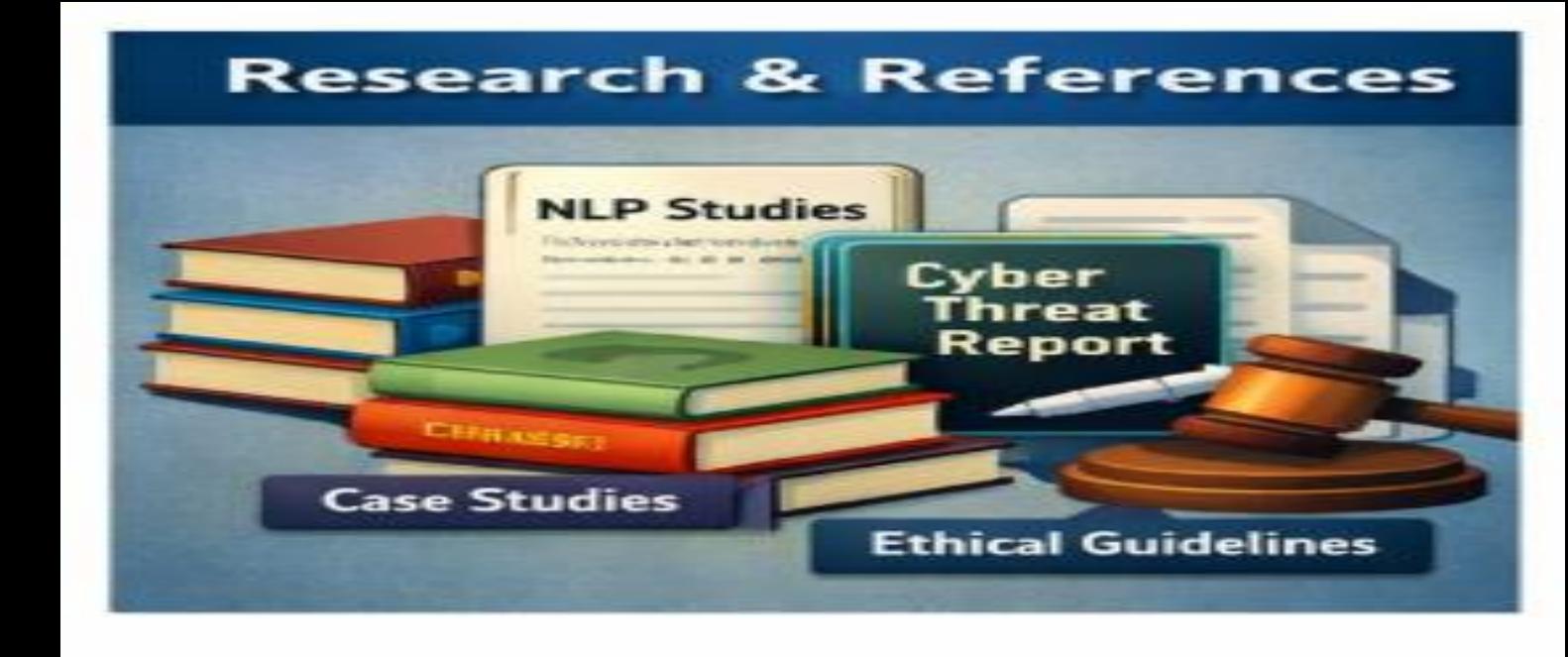
Challenges :

- Access restrictions or CAPTCHA on some forums.
- High volume of irrelevant or false data.
- Ensuring accuracy in NLP context detection.
- Maintaining data privacy and ethical scraping practices.

RESEARCH & REFERENCE

Research :

- Cyber Threat Intelligence (CTI) Studies - Learned how threat data is collected, analyzed, and used for early warning systems in cybersecurity.
- Natural Language Processing Techniques - Explored methods for text analysis, entity recognition, and context understanding in online discussions.
- Web Scraping Methodologies - Studied automated data extraction techniques and scheduling mechanisms for continuous monitoring.

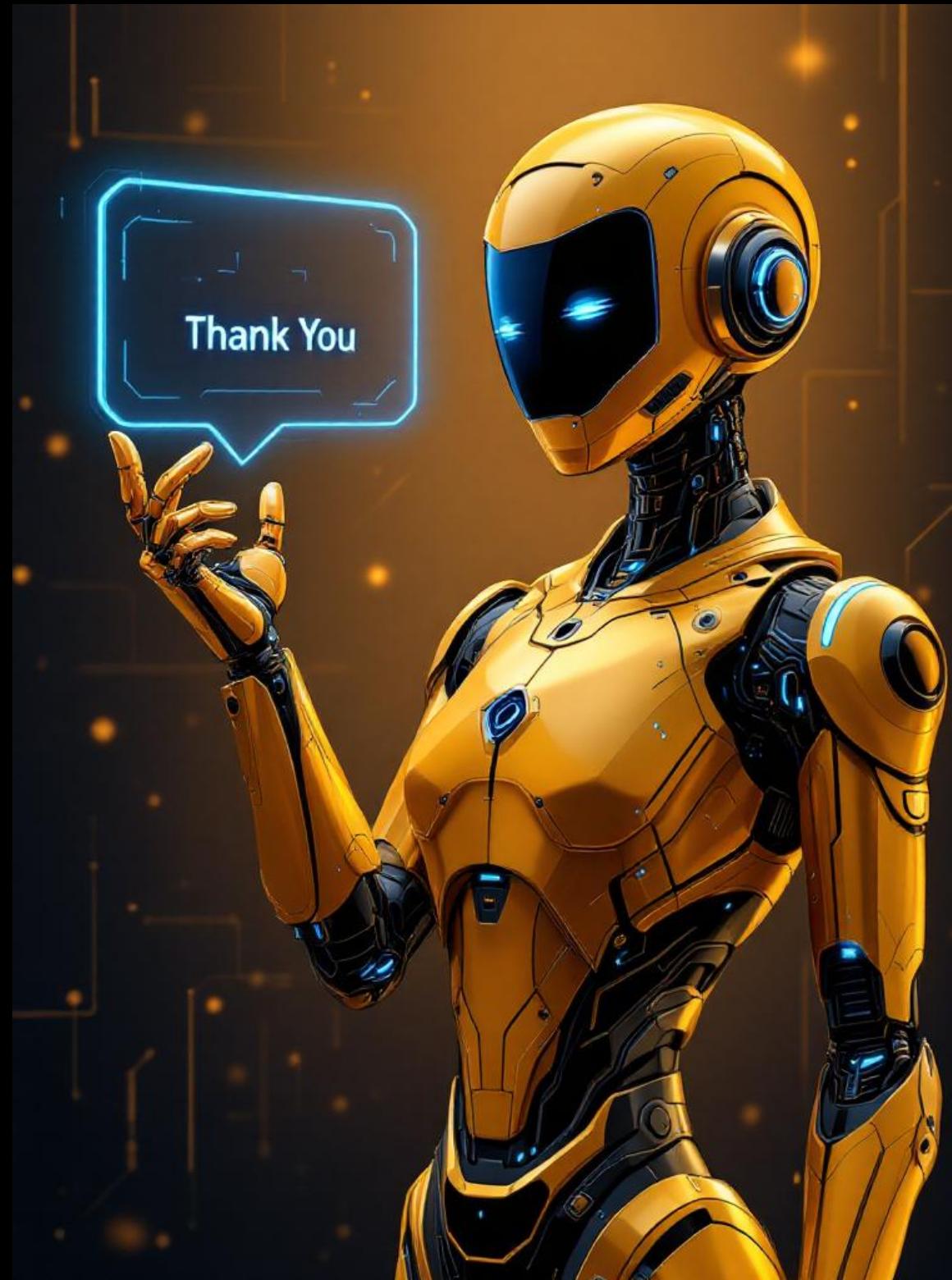


References :

- Research Papers on NLP and Cybersecurity - Used academic articles to understand advanced text-analysis and threat detection models.
- Open-Source Tool Documentation - Referred to official guides of scraping and database frameworks for technical implementation.
- Industry Security Reports - Examined annual cybersecurity reports and breach analysis documents for real-world insights.



HACK KRMU 5.0



THANK YOU