TITLE PAGE

- **Problem Statement ID – 1743**
- **Problem Statement Title - Parsing of Social Media Feeds**
- **Theme - Miscellaneous**
- **PS Category - Software**
- **Team ID - 26416**
- **Team Name - Lorven**

PARSING OF SOCIAL MEDIA FEEDS

SMART INDIA HACKATHON 2024

WHAT WE PROPOSE:

An innovative tool that **automates** the tedious process of scraping and filtering data from social media platforms, streamlining investigations by eliminating manual data collection efforts and reducing human error.

- **Automated actions:** Eliminates manual navigation, reducing human error during investigations using open-source **browser automation tools** like **Selenium** , **Playwright**.
- **Mobile Compatibility**: Solves platform-specific issues (e.g., Instagram, WhatsApp) with Appium for mobile behavior simulation.
- **Printable Reports**: Generates PDF reports with key screenshots, making evidence documentation and review easier using **ReportLab**.

WHY THIS APPROACH?

- Develop an automated tool to **extract and analyze social media data** (posts, timelines) across platforms like Facebook, Instagram etc.
- Automatically capture screenshots and **generate detailed reports** to assist investigators, enhancing accuracy and reducing human error.
- Ensure seamless functionality and **compatibility across platforms** to streamline investigative tasks.

What makes our solution unique?

Context-based Data Extraction:

❖This feature **narrows the search** to retrieve only the **relevant social media data** based on the case context provided by the agent.

REAL TIME MONITORING:

❖Our solution tracks the activities of **Potential Suspect** and generate **Alerts** for Investigator.

SENTIMENTAL ANALYSIS:

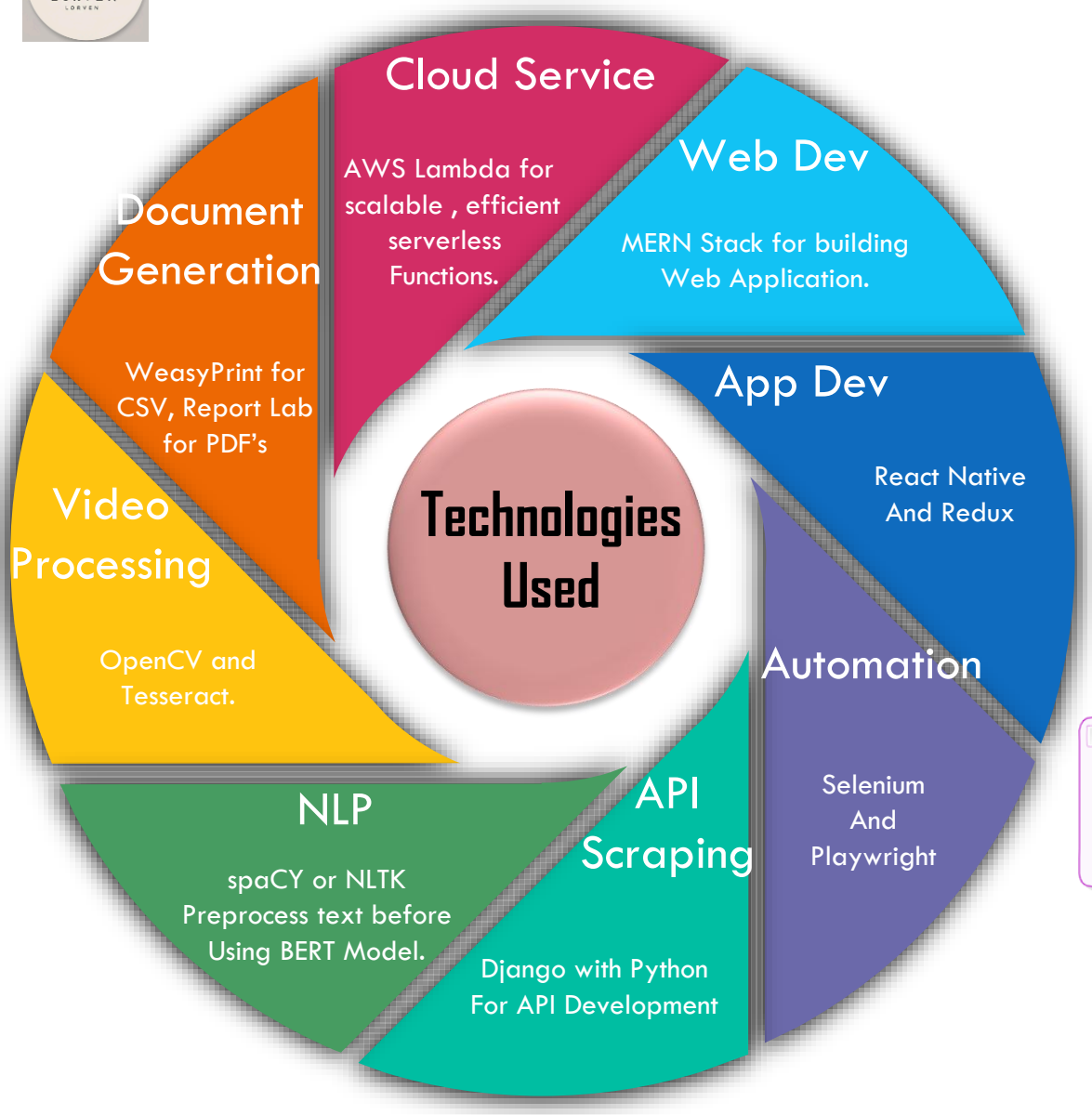
❖Unlike traditional tools, our solution integrates sentiment analysis to categorize the **Emotional tone** of posts and messages, providing investigators with **deeper insights** into the suspect's behavior.

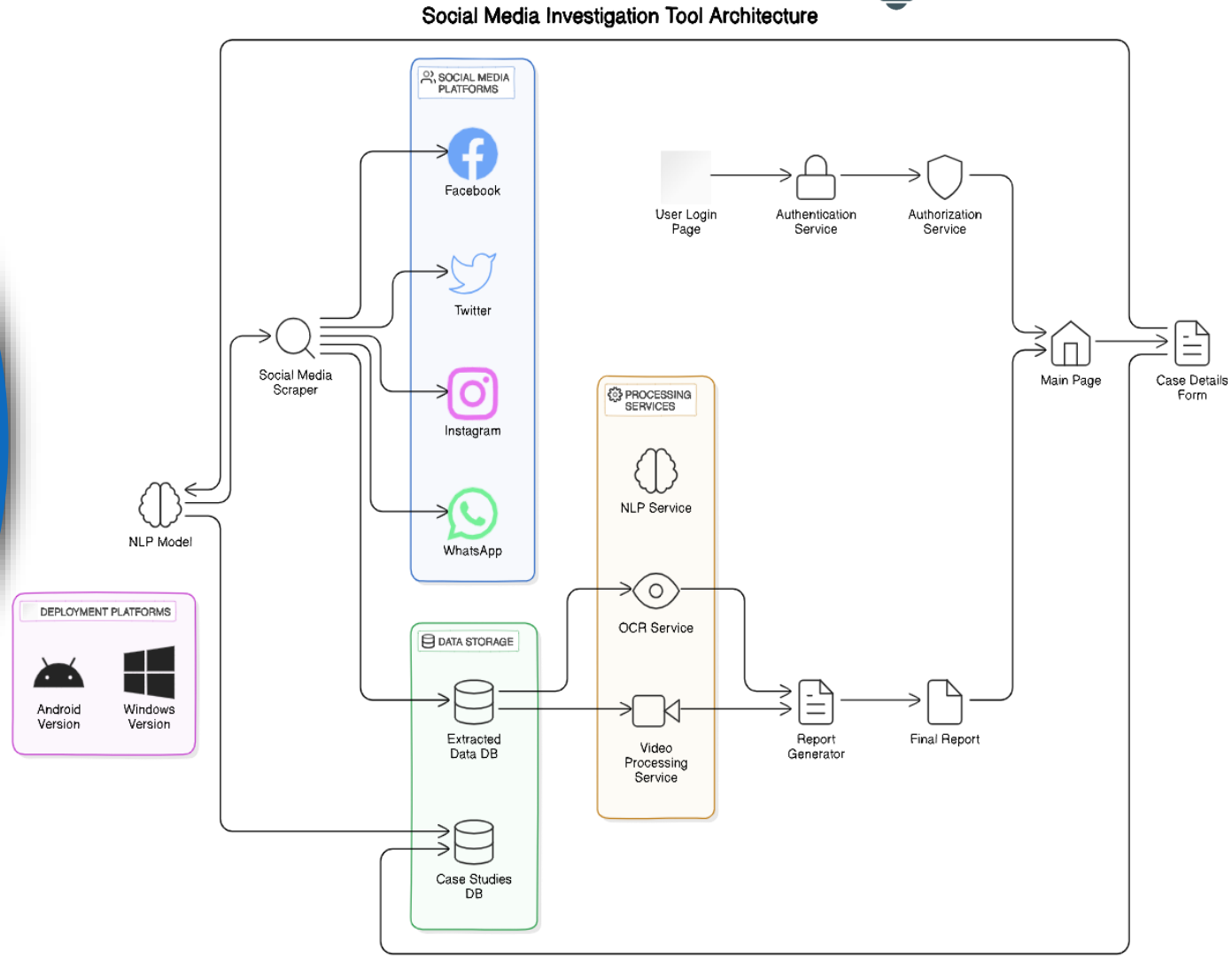
Video Message Processing and Analysis:

❖To process video evidence , video processing techniques can be used to extract relevant information such as spoken text, objects or scenes from the video.

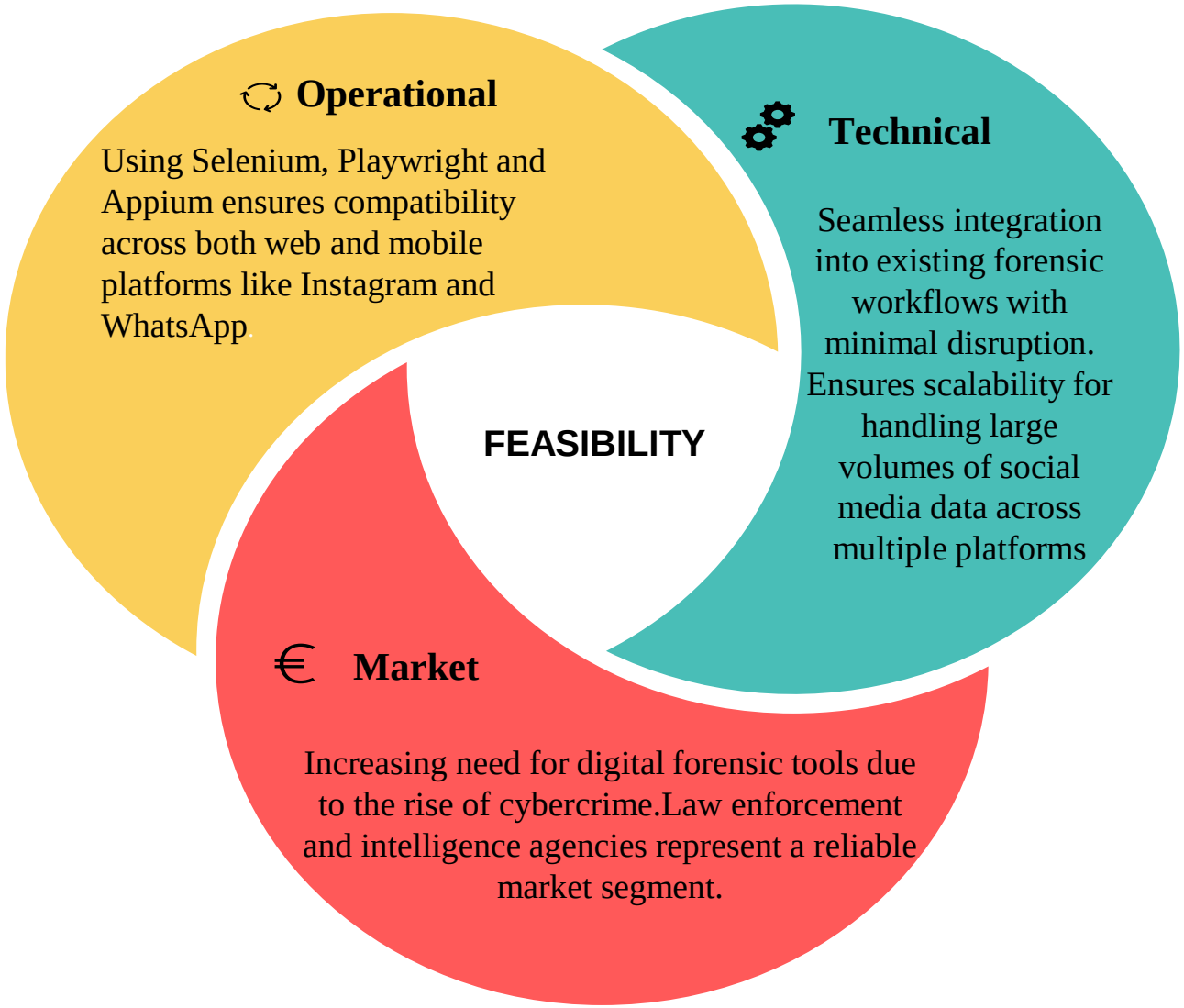
TECHNICAL APPROACH

SMART INDIA HACKATHON 2024





FEASIBILITY AND VIABILITY



Potential Challenges and Risks

•**CAPTCHA Issues**: CAPTCHA can disrupt real-time data extraction.
•Changes in platform architecture may break outdated data extraction methods.
•**Spam Detection:** Scraping may trigger spam detection, leading to IP bans or account restrictions or blocked access.

Strategies

•**CAPTCHA Issues**: Use manual intervention or external APIs to bypass CAPTCHA.
•Platform Updates: Regular updates are needed to handle changing interfaces. This may add to the long-term maintenance costs.
•**Spam Detection Mitigation:** Implement IP rotation, user-agent spoofing, and request throttling to avoid detection as spam. Additionally, use distributed scraping techniques across multiple proxies to minimize the risk of blocking.

IMPACT AND BENEFITS

Potential Impact on Target Audience:

Target Audience

National Investigation Agency **(NIA),** Police and Cybercrime Units, Private Security Firms, and other investigative bodies working in digital forensics and threat detection.

Law Enforcement Agencies

Facilitates rapid and precise extraction of social media data, optimizing the investigative process and speeding up critical decision-making.

Government Agencies

Strengthens efforts in safeguarding public security by providing a comprehensive tool for analyzing digital activities linked to criminal threats.

Private Security Firms

Assists in tracking and monitoring potential online risks, aiding private investigators in thorough background checks and security assessments.

Benefits of the Solution:

Deep Analytical Capabilities

- Leverages DL models and OCR capabilities to analyze the patterns in data and filter out the crucial information that is required for the investigation for context building and analyze all the information present in video and audio format to generate insightful reports for further investigation.

Live Data Tracking

- Monitors social media activities in real-time, offering timely updates for fast-moving cases.

Increased Productivity

- Automates the collection and evaluation of data, significantly cutting down investigation timelines.

Real-World Impact:

Streamlined Data Collection from Criminal Networks

- Efficiently retrieves information from platforms like Instagram and Facebook, allowing law enforcement to rapidly gather crucial evidence without worrying for Error of missing data.

Large-Scale Data Filtering

- Enables investigators to process vast amounts of social media content quickly, a task that would be unmanageable without automation.

RESEARCH AND REFERENCES

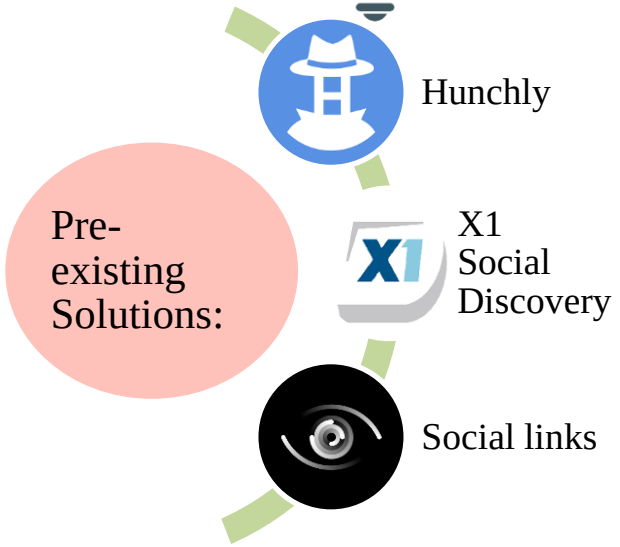
Reference Links:

- **ScienceDirect Article:** We referred to this article for guidance on building and training models based on social media data, especially focusing on practical techniques to improve model accuracy and robustness in real-world scenarios.

Source Link: https://www.sciencedirect.com/science/article/pii/S2772662222000273#tbl1

- **Scrapfly Blog:** This resource was utilized to understand the latest techniques for bypassing CAPTCHA during web scraping using browser automation tools, enabling smoother data collection from social media platforms.

Source Link: https://tinyurl.com/4jts55m4



Differences Between Our Tool and Existing Solutions:

Platform Support: Unlike X1 Social Discovery and Social Links, which mainly focus on desktop platforms, our tool offers dual support for both Android and Windows, allowing flexibility during investigations.

Automation and Screenshot Documentation: Unlike Hunchly, which is limited to browser-based data capture, our tool offers full automation for parsing and documenting social media content, significantly reducing human involvement. It also generates comprehensive sentiment analysis reports from the extracted data, helping investigators save time and gain valuable insights instantly. This enhanced functionality minimizes errors and optimizes the investigation process by providing deeper, real-time analytical results, which is not available in existing.

CAPTCHA Handling: Our tool integrates CAPTCHA bypass mechanisms, ensuring uninterrupted scraping of social media content, a feature lacking in many existing tools.Scalability and Real-Time Processing: Unlike most existing tools, our tool is designed to handle large-scale data extraction across multiple platforms efficiently.