SocialScan : Next-Gen Tools for Automated Social Media Evidence Collection

Synopsis submitted to

Shri Ramdeobaba College of Engineering & Management, Nagpur in partial fulfillment of requirement for the award of

degree of

Bachelor of Engineering

In

COMPUTER SCIENCE AND ENGINEERING

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

By

Mr. Anshul Jadhao

Mr. Arpit Agrawal

Mr. Lokendra Sinha

Mr. Raj Khatri

Mr. Sahil Dongare

Guide

Dr. Nisarg Gandhewar

Computer Science and Engineering (AI & ML)

Ramdeobaba University, Nagpur, Nagpur

440013

Problem Definition: Parsing of Social Media Feeds for Forensic Investigation

During forensic investigations, manually reviewing social media accounts can be error-prone and inefficient. An automated tool is needed to parse data from platforms like Facebook, Instagram, Telegram, and capturing posts, messages, timelines, friend lists, and account info. The tool should generate screenshots of this data in a documented format, allowing examiners to print or save relevant information. Additionally, as social media accounts often fail to load on desktops, the tool should have both Android and Windows versions to ensure accessibility and minimize human error.

Project Objectives:-

- Automate data extraction to streamline information gathering.
- Generate reports for efficient documentation and analysis.
- Ensure accurate data by analysis and store it in database.
- LLM fine- tuning for investigation purpose on scrapped data.
- Provide efficient behavioral investigated according to given prompt.

Proposed Plan of Work:-

Work	**No. of days required(estimated)**
Project Setup	**1 day**
Requirement Analysis and Feasibility Study	**2 days**
Technology Selection and System Design	**2 days**
Development of Social Media Extraction Modules	**3 days**
Data Processing and Storage	**3 days**
User Interface	**2 days**
Report Generation and Documentation Module	**3 days**
Testing, Validation and Security Enhancements	**3 days**
Deployment, Training, and Maintenance	**2 days**
Total	**21 days**

Methodology:-

- **Step 1: Input and Platform Selection**
  Create an interface for investigators to input social media details and select platforms for data extraction.

- **Step 2: Dynamic Content Handling**
  Use Selenium to handle dynamic content and scrape data from platforms with JavaScript or restricted access.

- **Step 3: Data Extraction and Parsing**
  Extract HTML with Selenium and parse using BeautifulSoup4 or lxml to isolate key data like posts and messages.

- **Step 4 : Data storing in Database**
  Store scraped data in database.

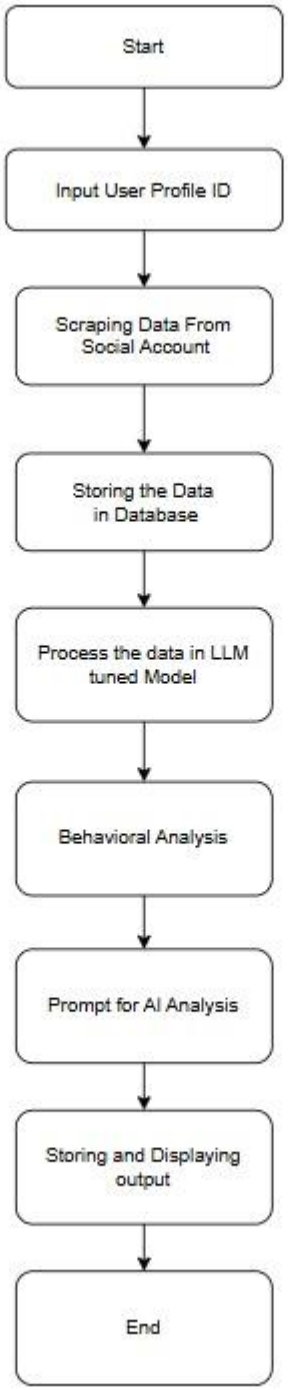
- **Step 5: AI-Powered Analysis**
  Integrate AI to analyze data, detect patterns, and flag suspicious behaviors or keywords.

- **Step 6: Fine-tuning on LLM Model**
  Fine-tuning an LLM for profile scraping improves structuring and behavioral analysis with greater accuracy and context.

- **Step 7: Security and Environment Management**
  Use python-dotenv to securely manage credentials and API keys, ensuring privacy and compliance.

- **Step 8: Testing and Validation**
  Test the tool across platforms to ensure accuracy, reliability, and adaptability to updates.

This Flow Diagram represent workflow of the project:-



Technology:-

- **Streamlit:** For building an interactive web interface.
- **LangChain:** For integrating language models and AI capabilities.
- **LangChain-Ollama:** For leveraging Ollama's language models within LangChain.
- **Selenium:** For handling dynamic content and JavaScript-heavy websites.
- **BeautifulSoup4:** For parsing HTML and XML documents.
- **html5lib:** For parsing HTML documents in a browser-like manner.
- **python-dotenv :** For managing environment variables and configurations.

Functional Specifications (Deliverables):

Web Application:

- A user-friendly Streamlit interface for URL input, scraping, and data display.

LLM fine-tuned model :

- Train model for profile investigation and behavioral analysis.
- Output according to prompt by AI analysis.

Project Scope:-

The project will focus on:

- Development of a scalable and robust AI social network web scraper for static and dynamic profiles.
- Collecting behavioral analysis using LLM for social background investigation.
- Investigated data of social media profiles.

Roll. No.	Name of Students	Name of Guide
21	Anshul Jadhao	Dr. Nisarg Gandhewar
22	Arpit Agrawal	
32	Lokendra Sinha	
45	Raj Khatri	
49	Sahil Dongare	

Approved by :

Dr. Nisarg Gandhwar
(Guide)