Social Media Parsing Tool for Digital Forensics: An Automated Approach to Evidence Collection and Documentation

by

Prince Raj
Ankit Das
Ayush Yadav
Isha Saha

Under the Supervision of
Alok Nath Pal
Assistant Professor, Department of CSE (Artificial Intelligence & Machine Learning)

Department of CSE (Artificial Intelligence & Machine Learning)
Narula Institute of Technology
(An Autonomous Institute)

Maulana Abul Kalam Azad University of Technology, West Bengal
(Formerly knows as West Bengal University of Technology)
December, 2025

Social Media Parsing Tool for Digital Forensics: An Automated Approach to Evidence Collection and Documentation

A Dissertation Submitted in partial fulfillment for the Degree of Bachelor of Technology (B.Tech), 5th Semester in CSE (Artificial Intelligence & Machine Learning)

by

Prince Raj(Roll No: 431023010134, Reg. No. 231270110438)
Ankit Das(Roll No: 431023010136, Reg. No. 231270110501)
Ayush Yadav(Roll No: 431023010127, Reg. No. 231270110379)
Isha Saha(Roll No: 431023020173, Reg. No. 231270110402)

Under the Supervision of

Alok Nath Pal

Assistant Professor, Dept. of CSE (Artificial Intelligence & Machine Learning)

Narula Institute of Technology
(An Autonomous Institute)

Maulana Abul Kalam Azad University of Technology, West Bengal
(Formerly known as West Bengal University of Technology)
December, 2025

Contents

CERTIFICATE OF ORIGINALITY

The Project entitled "Social Media Parsing Tool for Digital Forensics: An Automated Approach to Evidence Collection and Documentation" has been carried out by ourselves in partial fulfillment of the degree of Bachelor of Technology in CSE (Artificial Intelligence & Machine Learning) of Narula Institute of Technology, Agarpara, Kolkata under Maulana Abul Kalam Azad University of Technology during the academic year 2025-2026.

While developing this project, no unfair means of illegal copies of software etc. have been used and neither any part of this project nor any documentation have been submitted elsewhere or copied as far in our knowledge.

Signature of Prince Raj **Roll. No. 431023010134** **Reg. 231270110438**	**Signature of the Ankit Das** **Roll. No. 431023010136** **Reg. No.231270110501**
Signature of Ayush Yadav **Roll. No. 431023010127** **Reg. No.231270110379**	**Signature of Isha Saha** **Roll. No. 431023020173** **Reg. No.231270110402**

CERTIFICATE OF APPROVAL

This is to certify that the project entitled **"Social Media Parsing Tool for Digital Forensics: An Automated Approach to Evidence Collection and Documentation"** has been carried out by **Prince Raj, Ankit Das, Ayush Yadav, Isha Saha** under my supervision in partial fulfillment of the requirements for the degree of Bachelor of Technology (B.Tech.) in **CSE (Artificial Intelligence and Machine Learning** of Narula Institute of Technology, Agarpara, affiliated to Maulana Abul Kalam Azad University of Technology, during the academic year 2025–2026.

It is understood that by this approval, the undersigned does not necessarily endorse any of the statements made or opinions expressed herein, but approves the project only for the purpose for which it is submitted.

1. **Prince Raj** [University Roll No.: 431023020134, University Reg. No.: 231270110438]
2. **Ankit Das** [University Roll No.: 431023020136, University Reg. No.: 231270110501]
3. **Ayush Yadav** [University Roll No.: 431023020127, University Reg. No.: 231270110379]
4. **Isha Saha** [University Roll No.: 431023020173, University Reg. No.: 231270110402]

Signature of Supervisor	**Signature of HOD**
(Name of Supervisor)	**(Name of HOD)**

Abstract

Social media platforms have become vital repositories of digital evidence, yet law enforcement agencies face significant challenges in the manual collection, analysis, and preservation of this volatile data. This paper presents an automated Social Media Parsing Tool designed to streamline the extraction and documentation of digital evidence from platforms such as Facebook and Instagram. Developed as a Windows application using an Electron.js frontend and a Node.js backend, the system utilizes Selenium-WebDriver to mimic human browsing behavior, enabling the automated capture of posts, messages, timelines, and account information. To ensure data integrity and security, that is critical for legal admissibility, the tool leverages Google Firestore for encrypted storage and uses PDFKit to generate comprehensive, forensic-ready reports. By automating the data collection process, the tool minimizes human error, reduces investigation time, and maintains a strict chain of custody. Future enhancements include the integration of AI-based analysis and expansion to Android and web-based platforms to further support the evolving needs of digital forensics.

Chapter 1

Introduction

1.1 The Role of Social Media in Modern Investigations

In the contemporary digital landscape, social media platforms have evolved into a primary information source, fundamentally altering how communication occurs and how criminal inquiries are conducted. The vast quantities of data generated by users frequently contain critical evidence essential for resolving complex crimes, identifying potential suspects, and corroborating witness statements. Consequently, law enforcement agencies are increasingly turning to these digital platforms to gather real-time intelligence that provides deep insights into criminal behaviors, organizational networks, and ongoing activities.

1.2 Challenges in Digital Evidence Management

Despite its utility, the extraction of information from social media presents significant hurdles regarding the collection, analysis, and preservation of digital evidence. The sheer volume and velocity of social media data demand a high degree of accuracy, speed, and consistency; without these, vital evidence risks being lost or modified due to the highly dynamic and ephemeral nature of online content. To overcome these obstacles,

investigative bodies require specialized tools and technologies capable of streamlining data documentation while adhering to rigorous protocols for data protection and user privacy.

1.3 Proposed Solution: The Social Media Parsing Tool

The proposed Social Media Parsing Tool is designed to meet these logistical demands by offering an automated framework for the extraction, parsing, and precise recording of data across various platforms. Initially developed as a dedicated Windows application, the tool focuses on enhancing the efficiency of digital evidence gathering while drastically reducing the probability of human error. This automation ensures that investigations can proceed with both speed and forensic precision, ultimately facilitating quicker case resolutions.

1.4 Legal Integrity and Data Security

A core priority of this tool's architecture is the maintenance of data integrity to ensure that all collected evidence is legally defensible. By automating key collection steps and prioritizing secure storage, the tool ensures that the gathered data can withstand rigorous legal examination and be admitted as evidence in a court of law. This robust design not only optimizes the internal investigation process but also strengthens the overall evidentiary chain of custody.

Chapter 2

Problem Definition and Objective

2.1 Problem Definition

During investigation when the social media accounts of accused/suspect are opened for examination or creating Panchnamas, it would be better if some tool is designed which can automatically parse the data and provide the screenshot of the posts, messages, timeline, friend list, following, followers, account info, etc and provide screenshots in a documented form. * The examiner may choose to print the screenshots as per requirements. This will omit any human error during the process and also help to thoroughly reviewing the data found for the said social media account. * Separate options for Facebook, Twitter, Instagram, Telegram, WhatsApp, Google account etc may be provided in the tool. * Many a times, the social media accounts do not open in Desktops even if we have the right credentials and the examiner have to use a dummy android phone. So, two separate versions (Android and Windows) of this tool will be helpful.

2.2 Objective

The primary goal of the Social Media Parsing Tool is to fundamentally transform how law enforcement agencies handle digital evidence by enhancing the efficiency, accuracy,

and security of social media data extraction and documentation. To achieve this, the project focuses on the following detailed objectives:

1. Comprehensive Automation: The tool aims to replace manual labor with a fully automated system capable of extracting and documenting diverse social media data points, including user posts, private messages, account timelines, and detailed profile information from accounts logged into a Windows system. This automation is critical for reducing the high probability of human error during the evidence-gathering process.

2. Robust Data Security and Integrity: A central objective is the implementation of high-level security protocols to protect sensitive investigative data. By utilizing Google Firestore, the system ensures that all extracted information and screenshots are stored with strong encryption and strict access controls, thereby maintaining the chain of custody and preventing unauthorized tampering.

3. Operational Efficiency: The project prioritizes the optimization of backend logic to ensure that data-heavy tasks are completed rapidly without sacrificing the precision required for forensic work. This speed is essential to prevent the loss of volatile evidence due to the constantly changing nature of social media content.

4. Intuitive User Interface (UI) Design: The tool is designed to be accessible to investigators regardless of their technical background. Through a user-friendly Electron.js interface, the objective is to allow users to navigate the system, initiate complex extraction tasks, and generate professional reports with minimal training.

5. Scalability and Future-Proofing: While the initial release is optimized for Windows environments, a key objective is to build a modular architecture that supports seamless expansion. The design is intended to facilitate future development for other platforms, such as Android and Web-based interfaces, to adapt to the evolving needs of digital forensic teams.

Chapter 3

Scope of Project

3.1 Highlight Specific Aspects the Project Will Cover

The project aims to modernize the forensic workflow by addressing several critical functional requirements:

1. Comprehensive Data Automation: The system is designed to automate the extraction and documentation of a wide array of social media data, including user posts, private messages, account timelines, friend lists, followers, and general account information.

2. Forensic Evidence Documentation: A core objective is the generation of high-fidelity screenshots and comprehensive PDF reports that record digital evidence in a structured, documented form suitable for legal examination.

3. Data Integrity and Security: The project prioritizes the secure storage of all extracted data and images within Google Firestore, utilizing encryption and access control to prevent unauthorized access and ensure the integrity of the evidence.

4. Operational Optimization: By leveraging a Node.js backend and Selenium-WebDriver, the tool aims to perform these complex extraction tasks with high speed and precision, thereby minimizing the potential for human error and ensuring investigations

move forward accurately.

5. User-Centric Forensic Interface: The project focuses on creating an intuitive Electron.js interface that allows investigators to initiate tasks, track real-time progress, and access reports without requiring specialized technical expertise.

3.2 Boundaries and Focus of the Project

To ensure a reliable and deployable solution, the current scope of the project is defined by the following boundaries:

1. Primary Operating Environment: The initial focus is strictly on a Windows-based desktop application, as Windows is the predominant operating system utilized by law enforcement agencies.

2. Browser-Based Interaction: The tool's primary mechanism for data extraction involves interacting with social media profiles that are already logged into a web browser on the host system.

3. Platform Specificity: The tool provides dedicated, separate parsing options for a specific set of platforms, including Facebook, Instagram, Twitter (X), Telegram, WhatsApp, and Google accounts.

4. Future Scalability: While mobile support is identified as a critical need—particularly for instances where accounts cannot be accessed via desktop—the current architecture is designed for scalability to allow for future expansion into Android and web-based versions.

Chapter 4

Methodology

4.1 Data Collection

The initial phase of the methodology focuses on a comprehensive Requirement Analysis to bridge the gap between technical capabilities and investigative needs. This process involves identifying and analyzing the specific operational requirements of law enforcement agencies, particularly regarding the handling of digital evidence. Key data points to be targeted for collection include posts, messages, timelines, and account information from logged-in sessions. Furthermore, this stage defines the technical requirements for secure interaction with diverse social media platforms and the protocols necessary for storing sensitive data. We utilize an Agile development methodology, which facilitates iterative development and allows for the integration of regular user feedback to refine the data collection parameters.

4.2 Design

The System Design phase ensures the tool is built on a foundation that is modular, scalable, and secure. Detailed specifications are created for both the frontend and backend components to ensure seamless integration. The architecture emphasizes a modular de-

sign, allowing each part to operate independently—a critical feature for future-proofing the tool against evolving social media APIs.

- **Frontend Design**: Utilizing Electron.js, the interface is designed to provide a responsive, dashboard-driven experience for Windows users.
- **Backend Architecture**: The Node.js environment is selected to manage the complex logic required for high-speed data extraction and asynchronous processing.
- **Storage Strategy**: Google Firestore is integrated into the design to provide a cloud-based NoSQL solution that supports real-time synchronization and strong encryption.

4.3 Development

The Development phase translates the design specifications into a functional tool through three primary workstreams:

- **Backend Development**: Core logic is implemented using Node.js to handle data parsing, task scheduling, and communication with the database.
- **Frontend Development**: The user interface is built with web technologies (HTML, CSS, and JavaScript) within the Electron framework to ensure it is intuitive for investigators without technical expertise.
- **Automation Integration**: Selenium-WebDriver is integrated to simulate human-like interactions, allowing the tool to navigate complex social media profiles and capture high-fidelity screenshots automatically.
- **Reporting Module**: Developers utilize libraries like PDFKit to automate the generation of structured evidence reports containing embedded metadata and screenshots.

4.4 Testing and Validation

To ensure the tool meets the rigorous standards required for forensic work, a multi-layered Testing strategy is employed:

- **Unit Testing**: Individual components, such as specific data extraction functions, are tested in isolation using frameworks to ensure they function correctly as separate entities.
- **Integration Testing**: Interaction between the frontend, backend, and Firestore database is validated to ensure smooth data flow and a consistent user experience.
- **Automation Testing**: Selenium is utilized for end-to-end testing, simulating real-world browsing scenarios to validate the reliability of the automated extraction scripts.
- **Validation and Maintenance**: Following deployment, which includes packaging the app for Windows distribution , the tool enters a maintenance phase. This involves addressing bugs and updating automation scripts to stay compatible with social media platform modifications.

Chapter 5

Tools and Technologies

The Social Media Parsing Tool is built upon a high-performance technology stack selected to ensure forensic reliability, cross-process stability, and data integrity. The following sections detail the technical implementation of each architectural component.

5.1 Frontend Development: Electron.js Framework

The user interface is constructed using Electron.js, an open-source framework that enables the development of native desktop applications using standard web technologies: HTML, CSS, and JavaScript.

- Native Integration: Electron allows the tool to operate as a robust Windows application, providing the necessary low-level system access while maintaining a high-performance graphical interface.
- User Experience (UX): By utilizing web technologies, the interface remains responsive and intuitive, ensuring that personnel can initiate complex extraction tasks without extensive technical training.

5.2 Backend Infrastructure: Node.js Runtime

The core logic of the system is powered by Node.js, selected for its event-driven, non-blocking I/O model which is essential for handling intensive data operations.

- Asynchronous Processing: Node.js efficiently manages multiple asynchronous tasks, such as simultaneous browser automation and database write operations, without blocking the main execution thread.
- Task Management: The backend is responsible for orchestrating the entire lifecycle of an investigation, including user authentication, task scheduling, and error logging.
- Modular Architecture: The design follows a modular pattern, allowing for independent updates to the extraction logic or database handlers without disrupting the overall system stability.

5.3 Automation Engine: Selenium WebDriver

To achieve high-fidelity data extraction, the tool integrates Selenium WebDriver (JavaScript), a sophisticated browser automation suite.

- Human Mimicry: Selenium is configured to simulate human interactions—including clicks, scrolling, and navigation—enabling the tool to bypass simple interface barriers and access deep profile data.
- Dynamic Data Capture: The engine manages automated logins and navigates through complex social media timelines to capture real-time content and screenshots exactly as they appear to a human user.
- Forensic Reliability: By automating these actions, the tool ensures a consistent and repeatable extraction process, which is a fundamental requirement for digital forensics.

5.4 Data Storage and Security: Google Firestore

For the preservation of sensitive investigative data, the tool utilizes Google Firestore, a cloud-based NoSQL database.

- Security Protocols: Firestore provides enterprise-grade security, including automatic encryption of data both at rest and in transit.
- Real-time Synchronization: Its real-time data handling capabilities allow the frontend dashboard to reflect extraction progress and retrieved records instantaneously.
- Scalability and Access Control: The database is selected for its ability to handle large volumes of investigative data while enforcing strict, predefined access rules to protect against unauthorized entry.

5.5 Documentation and Reporting: PDFKit

The transition from raw data to admissible evidence is handled by the Reporting Module, which utilizes PDFKit.

- Automated Document Generation: PDFKit enables the programmatic creation of complex PDF documents without manual intervention, embedding screenshots and metadata directly into the report.
- Forensic Structuring: Reports are organized to highlight key evidentiary points such as timestamps, content types, and user interactions in a clear, coherent manner.

5.6 Testing and Quality Assurance

To validate the system's reliability, a comprehensive testing suite is employed:

- Frameworks: Mocha or Chai can be used for unit and integration testing to confirm the correct functioning of individual modules and their inter-component communication.

- Automation Validation: Selenium is repurposed during the testing phase to conduct end-to-end validation, ensuring the tool performs reliably under various network conditions and platform updates.

- Version Control: Git and GitHub are utilized for meticulous version management, tracking every change to the codebase and facilitating collaborative development.

Chapter 6

Expected Outcome

Decreased Investigation Time: The automation feature of our tool will reduce the amount of time investigators need to collect and record social media information. Automating the data extraction process enables investigators to analyze data more efficiently, potentially leading to faster case resolution. Enhanced Data Precision: Automation decreases the chance of human mistakes, guaranteeing precise and thorough data collection. The tool gathers all pertinent information immediately, reducing the chance of vital data being missed or inaccurately documented. Enhanced Security: Improved security measures are implemented when using Google Firestore for the Social Media Parsing Tool to securely store sensitive data with strict access controls. The tool focuses in security and privacy to safeguard data from unauthorized access without compromising its confidentiality. Scalability: Although the tool is initially intended for Windows, its modular structure will facilitate future expansion to other platforms. This ability to scale ensures that the tool can adjust to evolving requirements and stay current with the emergence of new platforms and technologies.

Chapter 7

Conclusion

An important change in the fields of online forensics and detection technology is the suggested Social Media Parsing Tool. With its help, investigators are free to focus on data analysis and valid conclusions made instead of the difficult process and subject to mistakes in gathering methods. The tool's ability and power are evident by using Google Firestore for storing data safely, Selenium for automatic processes, and an user interface made using Electron. Mainly focusing on flexibility and expansion, the tool's design makes it an excellent choice that can develop to meet future demands. Few of those demands consist of including various social media platforms, expanding to other operating systems, and adding AI-driven analysis. The tool not only meets expectations but also exceeds conditions needed for sensitive jobs. The tool's design emphasizes modularity and scalability, making it a robust solution that can evolve to meet future demands, such as supporting additional social media platforms, expanding to other operating systems, and incorporating AI-driven analysis. The strong focus on security and efficiency ensures that the tool not only meets but exceeds the standards required for investigative work. Through addressing the major challenges that law agencies nowadays face, the Social Media Parsing Tool has the potential to become an integral part of the digital investigator's toolkit, thus enhancing the speed, accuracy, and security of social media investigations. After proper implementation of the solution, the agencies will have access to a powerful, reliable, and secure tool that smoothens the process of social media data collection and documentation, which therefore will lead to more accurate and timely investigations.