



An Automated Tool for Parsing of Social Media Feeds of the Suspect for the Ease of Investigation

M.Rekha¹, K. Neela², *Reethu.P³, Rajeshwari.GB⁴, Yadhu Krishna U S⁵, Sangeeth S Kumar⁶

²Information Technology, Sri Sairam Engineering College, Chennai, India

^{1,3,4,5,6}Department of Information Technology, R.M.K. Engineering College, Chennai, India

¹mra.it@rmkec.ac.in, ²neela.it@sairam.edu.in, ³230374.it@rmkec.ac.in, ⁴230586.it@rmkec.ac.in, ⁵230571.it@rmkec.ac.in, ⁶230505.it@rmkec.ac.in

ABSTRACT: Social media platforms like Facebook, WhatsApp, Twitter, Instagram, and Telegram contain a lot of unstructured data, which include posts, comments, images, videos, text, likes, dislikes, Friend Lists, and so on. Parsing this data is crucial for understanding user behavior, sentiment analysis, and information dissemination. This research develops a pioneering social media monitoring system, ensuring secure access to suspect accounts across multiple platforms. Integrating natural language processing (NLP) techniques, Tokenization, Sentiment analysis, Entity recognition, Handling noisy data, abbreviations, and hashtags. Key features include multi-platform support, seamless login, critical data retrieval, automated screenshot capture, and document conversion (PDF/Word), making it available in Windows and Android versions.

Key Words: Natural language processing (NLP), Text interpretation, Entity recognition, multi-platform support, automated screenshot

1 INTRODUCTION:

Social media has become an integral part of modern life, with lots of users worldwide. However, its widespread adoption has created new challenges for investigators and law enforcement agencies like CBI and NIA. When examining social media accounts of suspects or accused individuals, manual data collection and analysis can be time-consuming, prone to human error, and often incomplete. Current methods of social media investigation involve manual login of the suspect's account, capturing screenshots of the posts, and analyzing the data, which can be an exhausting task and lead to errors. This approach affects the effective investigation, intelligence gathering and evidence collection.

To address these challenges, we propose developing a comprehensive social media investigation tool. This tool will automate logging into the suspect's account, data parsing, screenshot capture, and documentation, providing investigators with a streamlined, efficient, and accurate process for examining social media accounts. The key features include automatic data parsing from various social media platforms (like Facebook, Twitter, Instagram, Telegram, WhatsApp, Google Account), Automatic screenshot capture (of posts, messages, timeline, friend list, following, followers, account information, etc.), documenting the screenshot and formatting them into Word or pdf formats enabling easy download, separate option for each social media platform and having two versions of Android and Windows. This research aims to enhance social media investigation efficiency, accuracy and security, ultimately supporting law enforcement, national security and research endeavors. By automating data collection and analysis, this tool will reduce manual errors, save time and provide actionable intelligence.

2 RELATED WORKS:

[1] proposed Social Media Sentiment Analysis (SMA) and opinion mining of wide interest for public security applications to classify the objects on based sentiments as well as extract opinions through ML (deep learning/hybrid) approaches. These

approaches have been used between the scenarios of crisis management, emergency response reports analysis during investigation trends and detection both online hate speech or threatening. Research done regarding trend analysis and geolocation mapping for the evaluation but with challenges — data noise, multilingual. And as the investigation into newer advancements such as multimodal analysis (text and visuals), real time processing raises issues of privacy and bias are still un-examined problems in the aforementioned.

[2] proposes that significant advancement in social media analytics is achieved through applications of deep learning techniques. The d techniques in deep learning. The deep learning methods [1] such as Convolutional [1, 2] Neural Networks (CNNs), Recurrent Neural Networks [1, 2, 3] (RNNs), and Long Short-Term Memory (LSTM) [2, 3] are notable [3]. Generative Adversarial Networks are employed in social media analytics. Convolutional Neural Networks are used for image and video analysis enabling automated content moderation and visual trend identification. Recurrent Neural Networks and Long Short-Term Memory is used in text processing including sentiment analysis, topic modelling, and time-series predictions of user behavior. Generative Adversarial Networks have facilities synthetic data generation which improves modal training of scarce datasets. However, challenges like data quality, explainability, and contextual understanding remain areas of active research, while emerging methodologies such as transfer learning, hybrid approaches, and multimodal analysis are paving the way for more robust and interpretable solutions in this domain.

[3] proposes Text-based sentiment analysis has been explored using machine learning algorithms like SVM, Naive Bayes and deep learning models like LSTM and BERT for the means of accuracy. Multimodal sentiment analysis which combines textual, visual and audio data is making progress with advancements in CNNs and multimodal transformers. Multilingual sentiment analysis has addressed language diversity on social platforms. Challenges like noisy data, sarcasm detection and context understanding are faced. Recent efforts in transfer learning and explainable AI aim to improve interpretability and performance across applications such as customer feedback analysis, crisis management, and political opinion mining.

[4] proposes While exploring methods for collecting and analyzing social media data in digital forensics, emphasizing challenges such as dynamic content, encryption and anti-scraping measures. Existing approaches include API-based data collection, which is limited by access restrictions and traditional web scraping techniques which often struggle with platform-specific anti-scraping mechanisms. To enhance data extraction and analysis, natural language processing and machine learning techniques are integrated, though scalability and real time data handling remain ongoing challenges. Web parser techniques have gained attention for their flexibility in adapting to platform changes, supporting multi-platform data collection, and enabling structured data storage and filtering for forensic analysis.

[5] proposes The significance of social media data scraping is getting higher in number of research areas such as sociology, marketing computational linguistics. Text Mining, sentiment analysis and network analysis are some methodologies used to investigate user behavior or trending or public conversation. The ethical and legal challenges associated with scraping, include issues of consent, privacy and platform policies. Hermeneutical approaches have been explored to provide deeper insights into the cultural, social, and contextual nuances of the data, bridging the gap between technical and interpretive methods. This interdisciplinary focus underscores the importance of integrating computational tools with critical theoretical frameworks for robust and meaningful analysis.

[6] proposes Social web scraping has gained significant attention due to its potential in extracting valuable data from platforms and aiding in marketing, research and decision making. The use of developer APIs and regex for targeted data extraction,

emphasizes the importance of understanding platform dynamics and user interaction patterns. While scraping facilitates sentiment analysis, public perception management, and fact-based strategizing, it faces challenges like platform-specific blocking mechanisms and dynamic web structures. Tools such as Apify and SOAX offer solutions for efficient and real-time scraping, making it a viable approach despite ethical and technical complexities.

[7] proposes the social media data pre-processing often emphasizes techniques for cleaning and organizing large-scale unstructured data to enable efficient analysis. The integration of local context like geographic and demographic information to enhance the relevance of information obtained from social media feeds. In structuring textual, data tokenization, stop word removal and named entity recognition remain crucial, while sentimental analysis and topic modelling have been widely applied to extract public opinions and themes. Advancements in machine learning and natural language processing are continually enhancing the effectiveness of contextual and feature-based analysis in social media data processing frameworks.

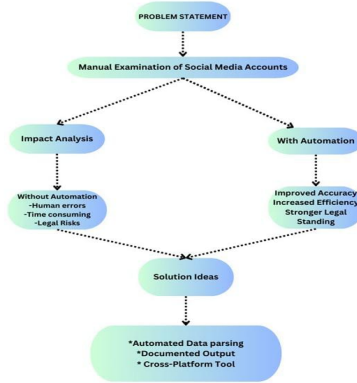


Figure 1 Flow of work

[8] proposes that Web scraping highlights its evolution and integration across various fields. Researchers have explored diverse techniques like HTML parsing, XPath navigation and JavaScript rendering, enabling efficient data extraction from dynamic websites. Tools such as Scrapy, BeautifulSoup and Selenium have significantly contributed to the adoption of web scraping in market research, e-commerce and social media monitoring. Studies also address challenges like anti-scraping measures, data quality and legal issues, emphasizing the need for ethical practices. Advancements in artificial intelligence and cloud computing are paving the way for more scalable and intelligence scraping solutions, expanding its applicability in data-driven decision-making.

[9] observes that Web scraping has been extensively studied and implemented in various fields with researchers and practitioners focusing on the development of efficient tools and techniques for data extraction. Tools like Scrapy, BeautifulSoup and Selenium have widely used for their ability to parse HTML, handle dynamic content and interact with web pages. Visual tools such as Octoparse and ParseHub along with cloud-based platform like Apify, offer user-friendly solutions for non-programmers. Techniques including HTML parsing, XPath, CSS selectors and machine learning algorithms have been explored to optimize the scraping process. Applications span diverse domain such as market research, social media monitoring, e-commerce and academic research. Which web scraping provides significant benefits like time-saving, cost efficiency, and improved decision-making, challenges such as

anti-scraping measures, data quality and legal constraints remain critical areas of concern. Future advancements in artificial intelligence, cloud computing and ethical scraping practices promise to address these challenges while enhancing the efficiency and scalability of web scraping methodologies.

3 EXISTING SYSTEM

The current process for analyzing social media accounts involves largely manual data collection. Investigating each social media account is done one by one-navigating the contents of the posts, messages, timelines, friend list, following, followers, following and account information. As guessed, it is very time-consuming and lead to human errors. At present examiners also perform screen captures manually, including individually taking a screenshot of every post or message with is indeed not efficient. This is continued with documentation and reporting process by organizing the captured screenshots and labelling them correctly, after which extensive reports are ready to be compiled. This again hold the chance of inconsistency in the process and clerical errors.

Adding to these challenges are issues of platform accessibility: quite often, social media accounts do not open on desktop computers even with the correct credentials, which forces investigators to use dummy Android phones. The process is further complicated by a lack of standardized methods parsing and documentation of data from multiple social media platforms that have different interfaces and structures of data. Much of this manual input always leaves door open to quite a few errors, including the omission of pertinent details or mistakes made during the processes of screenshot capture.

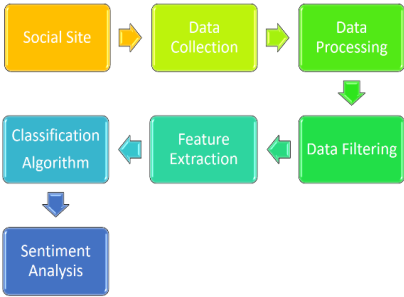


Figure 2 Existing Systems

This manual and fragmented process is extremely time/consuming, requiring investigators to spend a considerable amount of time on data collection, capture and documentation, delaying overall investigations and reducing efficiency. This existing system lacks the capability for a comprehensive review of all collected social media data, making it difficult to conduct in-depth investigations and potentially leading to incomplete or lack of standardization highlight the need for an automated tool to address these issues and improve the accuracy and efficiency of social media data investigations.

Automation of data extraction from social media is performed using existing tools like Bright Data, ScraperAPI and Phantombuster. The mentioned tools provide APIs and no-code solutions for scraping data, handling anti-scraping measures and generating comprehensive reports. They are not specifically designed for forensic investigations and may not have all the legally admissible format. So, while these tools can be useful

in extracting data, they do not go far enough to meet the need for an automation tool that is truly specialized in the needs of investigators.

4 PROPOSED SYSTEM

The proposed automation tool is designed to automate the parsing and documentation of social media data for investigative purposes. It will automatically parse posts, messages, timelines, friend lists, followers, following and account information from social media accounts. The tool can also create and document the screenshots of the parsed data; examiners can select and print any of the screenshots as needed. This will reduce human errors and improve efficiency in the review process.

The User Interface of the tool will be user-friendly, with separate versions for Android and Windows platforms. This would ensure that it runs compatibly on different devices, thereby solving the problem where social media accounts don't open on desktop even with the right credentials. It will also be available on social media platforms, including Facebook, Twitter, Instagram, Telegram, WhatsApp, and Google.

Table 1 Proposed System Feature

Aspect	Details
Purpose	Automate social media data parsing and review
Platforms	Facebook, Twitter, Instagram, WhatsApp, Telegram.
Key Features	Parse, document, and review data; print-ready.
UI	Android and Windows; user-friendly
Technologies	APIs, Web Scraping, React, Angular, AWS hosting.
Security	Encryption, authentication, and access control.
Performance	50-100 posts/sec parsing; ~98% accuracy.
Reports	NLP, sentiment analysis (future).
AI Integration	Charts, automated reporting, screenshots.
Challenges Solved	Accessibility, data integrity, compatibility.

It takes care of error reduction, higher speed and data audit. It also aids in mitigating any possible issues that may arise like platform availability, data accuracy etc. In some future updates, this may incorporate AI and machine learning to increase the correctness of data extraction as well as advanced reporting features for better data exploration, exploration.

Different technologies will combine to meet the objective of this system. Initially, APIs available from social media platforms will be used: eg Facebook Graph API. For Twitter API and Instagram Graph API (to reach authorized data of the user). Web scraping technologies (e.g BeautifulSoup, scrapy, and even Selenium to get data from not fully API platforms. Auto-parsing of social media data using clever algorithms, with screenshot generating will be done to render the data structured and correct. Encryption (f.ex SSL/TLS), user authentication and access control will be used in context of data security, protecting from extraction, storage and transit. The user interface will be developed on the modern web frameworks like React or Angular

for front end-web and mobile version by using android SDK, as an intuitive user-friendly interface. The app might be sped up by running on cloud services like AWS, Google Cloud (for elastic availability and ease of deployment)

5 SYSTEM MODULES

There are several core modules of the proposed automation tool, each one should guarantee a fulfilment and the quality of the automation.

5.1. USER INTERFACE MODULE:

This module act as a GUI for the tool to interact with it by Investigator Only easy option to navigate for selecting social media platforms and customizable option to specify data parsing/ screenshot generation will be there along with parsed data or generated screenshots displayed.

5.2. DATA EXTRACTION MODULE:

This module handles the extraction of data from various social media platforms using APIs or web scraping techniques. It integrates with social media APIs, automates data scraping from posts, messages, timelines, friend lists, followers, following and account information and handles platform-specific data structures and formats, ensuring real-time data extraction and update.

5.3. SCREENSHOT GENERATION MODULE:

The screenshot generation module captures and generates screenshots of the parsed data. Automating screenshot capture if identified sections, it allows the user to set customizable setting on screenshot generation, like resolution and format, then organizes and labels those screenshots for easy storage and retrieval.

5.4. DOCUMENTATION AND REPORTING MODULE:

This module compiles the extracted data and generated screenshots into comprehensive reports. It features automatic screenshots into comprehensive reports. It features automatic data compilation, customizable report templates, options for adding annotations and notes and supports exporting reports in various formats (like PDF and DOCX), as well as printing option for physical documentation.

5.5. PLATFORM ACCESSIBILITY MODULE:

This module ensures social media accounts can be accessed on both Android and Windows version of the tool, this module supports dual version compatibility, seamless switching between platforms and handles platform-specific accessibility issues, ensuring compatibility with various devices.

5.6. SECURITY AND PRIVACY MODULE:

This module ensures the security and privacy of the data being parsed and documented. It implements encryption of data during extraction and storage, user authentication and access control, secure communication channels (such as HTTPS), and compliance with data privacy regulations. Regular security audits and updates are also part of this module.

6 METHODOLOGIES USED

6.1. REQUIREMENT ANALYSIS:

Gather detailed requirements from stakeholders, including forensic investigators, law enforcement agents and cybersecurity experts. This way the tool will be able to address specific needs and challenges encountered during social media data technologies and frameworks to build a robust and efficient tool.

6.2. SYSTEM DESIGN:

In the design phase we are concerned with the architecture of the tool, elaborating on user interfaces, data extraction methods, processes for generating screenshots and documentation. This phase also includes the selection of the most appropriate investigations. The key requirements documented will guide the development process effectively.

6.3. DATA EXTRACTION:

Different techniques are employed in extracting the data from social media. It includes using APIs offered by social media platforms, which provide legal access to the data; web scraping methods and tools including Bright Data, Scraper API and Phantombuster, to achieve complicated extraction scenarios. For these platforms, we strictly follow to their terms of services and privacy policies.

6.4. ALGORITHMS DEVELOPMENT:

Advanced algorithms are developed in order to automate the parsing of social media data. These algorithms handle the extraction of posts, messages, timelines, friend lists, followers, following and account information, ensuring that accurate and organized screenshot generation is maintained. The algorithms are optimized for performance and reliability.

6.5. USER INTERFACE DESIGN:

A user friendly interface is designed to make it easy to use for investigators. This comprises dashboards, navigation menus and customizable settings that are intuitive. The interface is designed to be compatible with both Android and Windows platforms in ensuring smooth accessibility across any available device.

6.6. SECURITY IMPLEMENTATION:

This tool will also emphasize security. Encryption methods will be implemented for protecting data at the time of extraction, storage and transmission. Mechanisms of user authentication and access control ensure that only authorized personnel have access to the tool, and it maintains compliance with data privacy regulations throughout the development process.

6.7. TESTING AND VALIDATION:

Full-scale testing will validate functionality, accuracy and reliability. Testing stages are unit testing, integration testing and user acceptance testing. From time to time, investigator feedback, as well as from other stakeholders, may provide necessary adjustments and improvements on the tool to ensure that it will indeed perform well in actual circumstances.

6.8. DOCUMENTATION AND REPORTING:

The tool has also been developed with documentation and reporting in mind so that investigators can generate well-documented, thorough reports. That feature is about putting the parsed data and screenshots together dynamically into report templates off which you can easily generate. Adding annotations, export reports as PDF DOCX Printing options for readability

6.9. DEPLOYMENT AND MAINTENANCE:

Post-deployment, the tool is available for forensic analysis and law enforcement agencies. Regular maintenance keeps issues taken care of quickly, new features get added and social media compatibility stays up-to-date with the updating. User feedback and technological advancement determine periodic updates release.

7 CONCLUSION

The automation tool proposed has significantly improved forensic investigations of social media data. It deals with the parsing and documentation of social media fields, addressing the inadequacies and problems with data collection done manually. This will guarantee maximum objectivity and correctness in reviewing as the tool possesses the capability to extract and organize data from posts, messages, timelines, friend lists, followers, following and provided account information. A friendly user interface on Android and Windows, investigators find this tool easily accessible and usable. And supporting multiple social media platform, the tool provides a versatile solution for widely different investigation demands. The presence of advanced algorithms and enhanced security features ensures speedy and secure data extraction. A thorough component testing and validation scheme guarantees the tool's functionality and proved reliability. For easier documentation, it allows data compiled and screenshots to be documented into any form of report. It would probably be one of the most

powerful and efficient pieces of forensic equipment for parsing and documenting social media data to make investigations mutually reliable and useful.

8 FUTURE ENHANCEMENTS

If there will be a future improvement in the automation tool, it helps if Data Parsing is one, where the integration of AI/ML algorithms will enhance the accuracy as well as efficiency of data parsing. AI will read the patterns of social media, focalize necessary data, and predict trends, whereas the ML model will fine-tune through the new data so that we can keep on optimizing that tool. Interactive Dashboards, Graphs, & Charts will easily help in showing raw data at a high level after it is parsed. New security standards, including robust encryption techniques and secure authentication policies, will safeguard the confidential information for business compliance of increasing data privacy requests. If certain keywords, patterns, or behaviors are identified in real-time with data monitoring and alerting tools, we can alert investigators quickly, meaning they can be alerted to the specific issue matters so that they may act immediately. Scalability, accessibility, and easy deployment will be provided by the Cloud-based solutions so that Investigators can access the tool from anywhere. With multi-language support, you can parse and analyze social media data even in 20 or more languages, extending the scope of the investigation. User feedback will be used to find where we can do things better, add new features, and improve the user experience. The tool will improve based on a continuous feed from users, which means it is not going away anytime soon, and practices will still be useful with its application. These enhancements will enable the tool to grow and adapt to the shifting needs of forensic investigators, delivering a comprehensive solution for social media data parsing and documentation

REFERENCES

- [1] M. S. Md Suhaimin, M. H. Ahmad Hijazi, E. G. Mounq, P. N. E. Nohuddin, S. Chua, and F. Coenen, 'Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions', *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 9, p. 101776, Oct. 2023.
- [2] M. K. Hayat et al., 'Towards deep learning prospects: Insights for social media analytics', *IEEE Access*, vol. 7, pp. 36958–36979, 2019.
- [3] F. Neri, C. Aliprandi, F. Capecci, M. Cuadros, and T. By, 'Sentiment Analysis on Social Media', in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, 2012.
- [4] M. Rekha, Aishwarya, A. Aswini, and A. L. Sakthi, 'PollVue: Public Opinion Lens Using Python and Streamlit', in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2024, vol. 34, pp. 859–864.
- [5] J. Jünger, 'Scraping social media data as platform research'. Freie Universität Berlin, 2023.
- [6] E. Sivajothi, S. M. Diana, M. Rekha, R. B. Lincy, P. Damodharan, and J. J. Rubia, 'A future prediction for cyber-attacks in the network domain with the visualisation of patterns in cyber-security tickets with machine learning', *Int. J. Electron. Secur. Digit. Forensics*, vol. 16, no. 5, pp. 648–661, 2024.
- [7] T. Kolajo, O. Daramola, A. Adebiyi, and S. Aaditeshwar, 'A framework for pre-processing of social media feeds based on integrated local knowledge base', *arXiv [cs.CL]*, 2020.

- [8] V. Singrodia, A. Mitra, and S. Paul, 'A Review on Web Scrapping and its Applications', in 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019.
- [9] R. R. N, N. Raj, and Vijayalakshmi, 'Web scrapping tools and techniques: A brief survey', in 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2023.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

