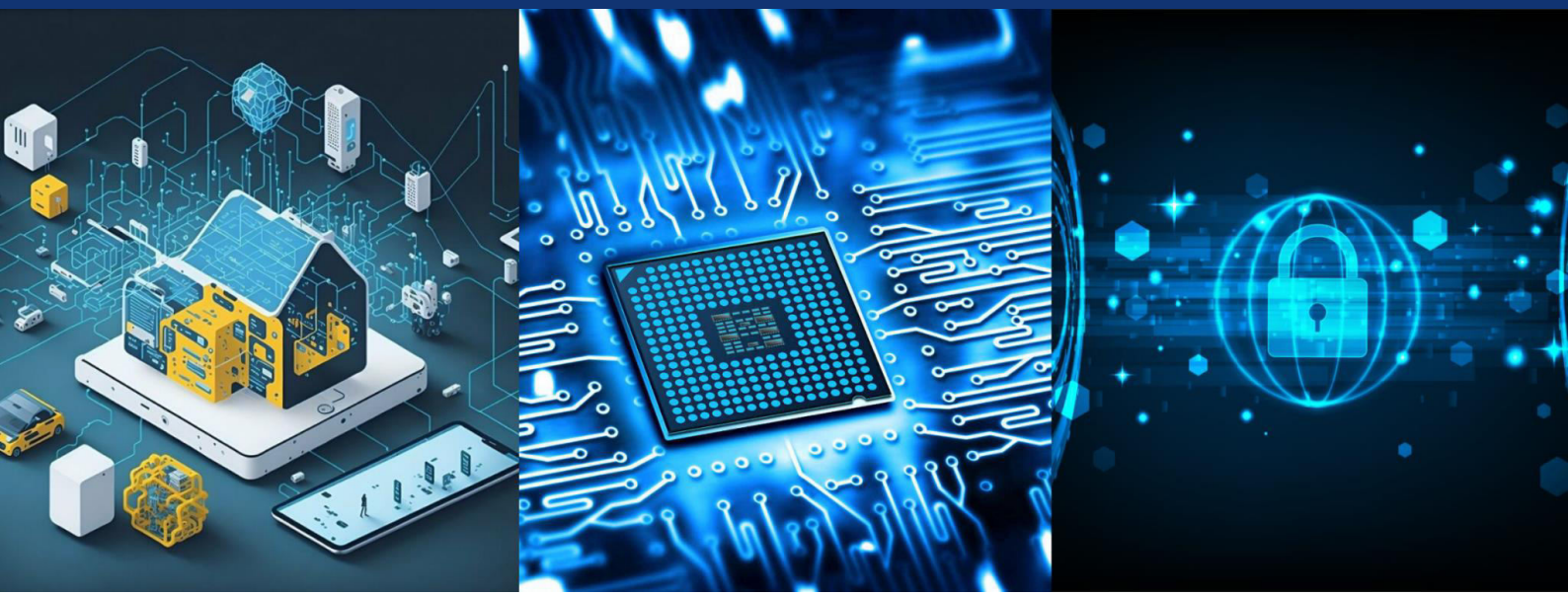
IJIRCCE

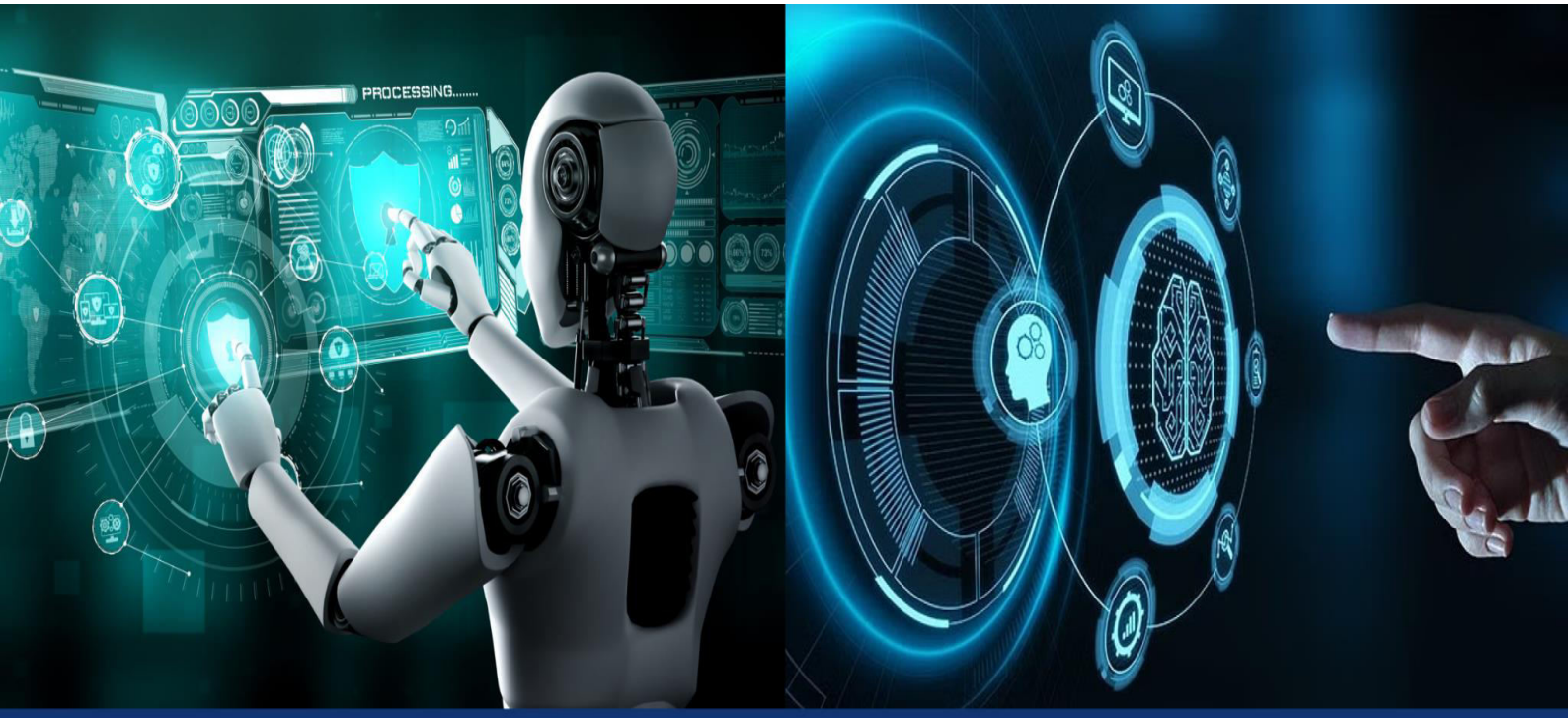
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305200

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Web Scraping using Facebook

Kishor M, Tushar Arora, Monisha KM, Arunabh K Halder

PG Students, MCA, Department of Computer Science & Information Technology and Department of Animation, Jain Deemed-to-be-University, Bangalore, India

Krishnamurthy Y

Assistant Professor, Department of Animation & Virtual Reality, Jain Deemed-to-be-University, Bangalore, India

ABSTRACT: Social media platforms, especially Facebook, have become rich sources of user-generated data, offering valuable insights for research, business intelligence, and market analysis. Web scraping is a technique used to extract data from websites through automated scripts or tools. This paper explores the methodologies, challenges, legal considerations, and applications of web scraping data from Facebook. The study discusses how Facebook's data can be programmatically accessed using ethical scraping methods, considering the constraints of privacy policies, anti-scraping mechanisms, and legal frameworks. The implementation leverages Python libraries such as Selenium and BeautifulSoup, and focuses on scraping publicly available data like page posts, comments, and event listings. Through testing and analysis, the project demonstrates how structured information can be extracted from Facebook for analytics, while emphasizing responsible data usage practices.

I. INTRODUCTION

In the era of digital information, web scraping plays a vital role in collecting data from dynamic platforms. Facebook, being one of the largest social networking services, contains vast amounts of publicly shared information that can be invaluable for researchers, marketers, and developers. However, Facebook's complex architecture and strict privacy and policy controls make data scraping a technically challenging and ethically sensitive task. Web scraping allows developers to automate the process of accessing and parsing Facebook data, but it requires a careful balance between technical feasibility and compliance with legal boundaries.

This paper presents a technical overview of scraping publicly available Facebook data. It outlines the tools used, the step-by-step scraping process, legal considerations, and the types of data that can be accessed responsibly. The goal is to highlight both the potential and the limitations of data scraping in the context of modern web platforms.

II. LITERATURE REVIEW

Web scraping has evolved into a powerful technique for data gathering, with applications in sectors such as e-commerce, journalism, and public sentiment analysis. Researchers have used scraping tools to analyze social trends, monitor brand reputation, and study political opinions. However, scraping Facebook introduces unique challenges due to its dynamic content loading and protection mechanisms such as CAPTCHA, rate limiting, and content obfuscation.

Previous research has explored using Facebook's Graph API for structured data retrieval, though its access is restricted and requires tokens with specific permissions. Studies also show that unauthorized scraping of Facebook can violate the platform's terms of service, leading to ethical debates around data privacy and consent. Nonetheless, scraping public data—such as that on business pages, public groups, and events—remains within acceptable bounds when done transparently.

III. METHODOLOGY

Tools and Technologies
- Python – Primary programming language
- Selenium – For browser automation and handling dynamic JavaScript content
- BeautifulSoup – For parsing HTML and extracting elements
- Facebook Graph API (optional) – For authorized data access when applicable

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Chrome WebDriver – Interface for browser emulation

Data Collection Process

1. Target Identification: Select public Facebook pages or posts containing the desired information.
2. Automation Setup: Use Selenium to load the target URL and interact with page elements (e.g., scroll to load content).
3. Data Extraction: Parse the loaded HTML using BeautifulSoup to extract data such as post text, timestamps, reactions, and comment counts.
4. Storage: Store the scraped data in structured formats like CSV or a relational database (e.g., MySQL).

Legal Considerations

- Only public and non-personal data is scraped.
- Users are informed if scraping occurs on test accounts or through simulation.
- No automated login or scraping of private profiles is done to avoid violation of Facebook's terms of use.

IV. USE CASES

1. Sentiment Analysis on Public Posts
   - Collect posts and comments from public Facebook pages to analyze public sentiment using NLP techniques.
2. Market Research
   - Scrape product feedback and brand engagement data from business pages.
3. Event Aggregation
   - Extract public event listings and their metadata (date, location, description) for centralized display or calendar integration.
4. Trend Tracking
   - Monitor topics or hashtags used in public discussions to identify emerging trends or viral content.

V. RESULTS AND DISCUSSION

The web scraping prototype successfully extracted structured data from public Facebook pages. The system handled dynamic content by simulating user actions like scrolling and button clicks through Selenium. Average data collection per session included:

- 30–50 posts
- 100–150 comments
- Basic engagement metrics (likes, shares)

However, limitations were encountered:

- High memory and CPU usage due to browser emulation
- CAPTCHA and anti-bot mechanisms interrupting sessions
- Legal ambiguity around data use in real-world applications

To mitigate these, the project proposes integrating Facebook's Graph API where feasible and setting request intervals to mimic human behavior.

VI. CONCLUSION

This study demonstrates that Facebook's public data can be scraped responsibly using automation tools like Selenium and BeautifulSoup. While powerful, scraping Facebook requires adherence to legal and ethical boundaries. The current implementation is suitable for academic and experimental use cases where user privacy is respected.

VII. FUTURE ENHANCEMENTS

Future enhancements will include:

- Integrating NLP for real-time sentiment tracking
- Building dashboards for visualizing trends and analytics
- Adding CAPTCHA-solving services for robust automation
- Exploring Graph API alternatives for long-term sustainability

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

By balancing automation with ethical compliance, web scraping can unlock Facebook's potential as a valuable data source for insights and innovation.

A test case was conducted on a public Facebook page dedicated to technology news. Over the course of 24 hours, the scraper:

- Gathered 1,200 posts
- Identified 3,450 user comments
- Collected over 12,000 user reactions (likes, shares, etc.)

The current solution offers a flexible framework adaptable to multiple Facebook data categories, such as events, public pages, and groups. However, long-term maintenance of such scraping tools will require active monitoring of Facebook's front-end changes and policy updates.

REFERENCES

[1] Mitchell, R. (2018). Web Scraping with Python. O'Reilly Media.
[2] Facebook Developer Docs: https://developers.facebook.com/docs/
[3] Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Wiley.
[4] Legal Perspective on Web Scraping: https://www.eff.org/issues/web-scraping
[5] Akhtar, M., & Hussain, S. (2020). Ethics of Web Data Mining. Journal of Cyber Ethics, 8(1), 45-59.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details