Research on Twitter Scraping

Automated process of collecting data from twitter such as tweets , user Id's, hashtags etc..

there are automated tools to integrate this process.. to gain insights, metrics validation, new trends, user's behaviour and market sentiments.

file format→ CSV or JSON, we can do further analysis from these formatted files.

How scrapping works?

→ using automated tools (to retrieve required data from the platform) they are bots or maybe a software. that sends request to retrieve the required information.

→ extract data (tweets- the content of the posts, including replies, retweets, etc.

→user profiles: bio, follower counts, usernames, etc.

→ hashtags : data and content under specific hashtag for trend analysis

→ engagement metric: likes, views and comments on the particular posts

we can structure the output in spreadsheets(CSV or JSON)- easy to organize and analyze

Scraping methods:

1. Twitter APIs : these APIs provide access to Twitter's data
2. Web scrapping tools: software/browser extensions to simplify the process without code
3. Py libraries: snscrape and Tweepy (data extraction using python libraries)

Is scraping data legal? yes, it is legal to scrape publicly available data

things to avoid:

→ avoid scraping private data without explicit consent

→don’t overload the server

→transparency about the source of information

Scraping Twitter(X)

→Latest & recommended version: X API v2

→ legal risks, privacy policy changes, paid API restrictions (due to request limits)

→ API turns into error if the limits are exceeded

How the platform detects scraping?

1. unusual requests rate limit responses
2. IP blocking :Multiple requests from one IP cause per-IP blocklists
3. Account fingerprinting : repeated automated request from same account
4. CAPTCHA’S and bot detection: detects non human usage by giving JS challenge.

Reasons for IP/ ID blocks:

Running large scale automated scraping using logged-in accounts (sudden spikes in activity).

Using outdated or reverse engineered private endpoints (GraphQL & undocumented APIs) that the platform can detect.

Excessive page loads, parallel connections, or scraping pages that require JS (which forces headless browsers and is detectable).

Alternatives:

Official X , Twitter APIs (main choice)

Data resellers(choose who provide legal and legitimate data, might be costly)

Open Source tools, sncrape → widely used for research purposes

sncrape: no code required

sncrape can scrape twitter data using the python library “snscrape” to easily pull millions of historic tweets and save them off on the computer. This can be used to create data for analysis or just archive off tweets quick and easy.

TWEET SCOUT→ crypto tool research

third party platform that surfaces twitter(X) signals for crypto and investment researches

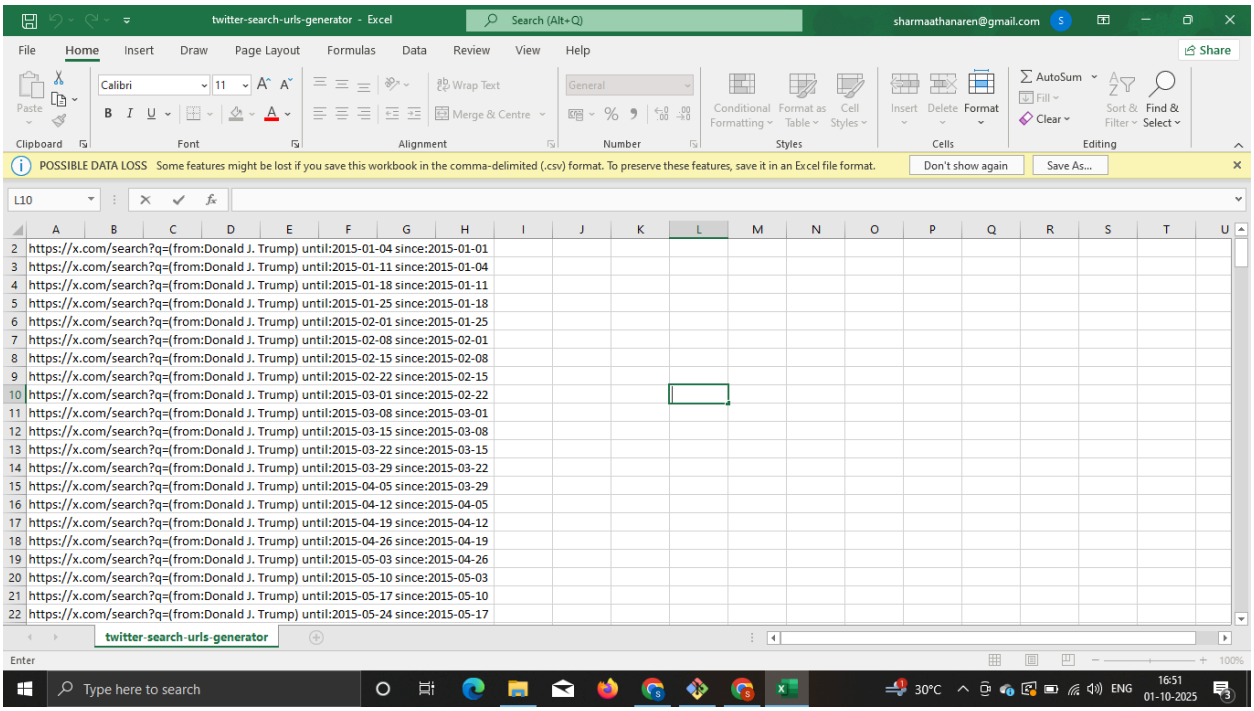
short note: More than $1.5 billion has been stolen by cryptocurrency and NFT scammers during the last years because there is no working tool that can show the entire list of foundations, influencers, and projects interested in an exact account, the number of bots in followers, etc. There are no clear indicators of the trustworthiness of crypto projects on social networks. TweetScout solves all these problems by bringing the entire crypto Twitter into one powerful analytical tool.

Scrape tweets, post, comments
X data such as posts, replies, and hashtags is typically collected either through API or using scraping tools. The official API provides structured tweet objects containing text, author details, engagement metrics, and entities like hashtags, along with identifiers .

Researchers can query user timelines, specific hashtags, or conversation threads, with results paginated and subject to rate limits.

Community tools like sncrape parses the same JSON data the web interface uses, allowing extraction of tweets, comments, and hashtag streams without API keys, though these methods.

I used a local scrapping tool(lobst.io) to see how the posts of a partical person gets scrapped with timeline



from Donald J. trump → filter tweets only from that account.

since: 2015-01-01→ start date (inclusive).

until: 2025-01-04 → end date (exclusive, i.e., tweets before this date)

End solution:

Define goal(what you need to scrape) —→ collect the tweets—→ combine the files—→ clean and preprocess —→ do required analysis —→ store and present