[1] Target Selection

↓

Decide WHAT to scrape

(e.g., forum posts, news articles, product reviews)

[2] Inspect & Identify Selectors

↓

Open DevTools → Inspect Element

Find stable CSS selectors / attributes for your target data

[3] Fetch Page Content

↓

├─ Static Page → requests.get(URL)

└─ Dynamic Page → Playwright/Selenium to render JS

[4] Parse HTML / JSON

↓

Use BeautifulSoup / lxml to navigate DOM

Extract text, links, images, or metadata

[5] Clean & Structure Data

↓

- Remove HTML tags

- Normalize whitespace

- Store in Python list/dict

[6] Save Data

↓

- JSON (structured, easy for NLP)

- CSV (quick sharing)

- In-memory (if processing immediately)

[7] NLP Processing

↓

- Tokenization

- POS tagging

- Text classification (e.g., harmful / safe)

- Sentiment or keyword extraction

[8] Output / Visualization

↓

- Save processed results

- Generate summary stats

- Display in dashboard or console