Collecting Social Media Data

Tools for Obtaining Data from Social Media Platforms

Annika Deubel, Johannes Breuer, Katrin Weller

Social media platforms provide an internet-based form of communication for users to have conversations, share information, and create web content. The vast amounts of data produced by the increasing integration of social media into people’s daily lives is not only of interest to platform providers, but also to researchers in various scientific fields of digitalization research. The data from social media platforms can be used to gain insights into various behavioral, usage, and interaction patterns.

The analysis of this data has become increasingly popular in fields such as sociology, political science, linguistics, and computer science. For example, it allows for understanding collective sentiment, user behavior during global events, monitoring public opinion, identifying key topics in public discussions, and detecting popular routes followed by social media users.

Compared to other types of data that are commonly used in the social and behavioral sciences, such as survey or interview data, social media data are non-designed, i.e., they occur naturally and in large quantities. Researchers can, hence, harvest a massive amount and diverse range of user-generated content and other digital trace data without the need for data collection procedures that rely on self-reports. Once equipped with the resources for accessing and acquiring data, researchers can potentially scale data harvesting without expending a great deal of resources and often with less lead time compared, e.g., to survey research.

Introduction

In general, social media data can be obtained via three main options. Firstly, researchers can collect the data themselves, typically either through web scraping or the use of application programming interfaces (APIs). Secondly, privileged access can be gained by cooperating with companies that produce or hold these data. As a third option, the data can be purchased from market research companies or data resellers (Breuer et al., 2020).

Among all options, collecting data through APIs or web scraping have been the most widespread among researchers working with social media data (Acker & Kreisberg, 2020; Breuer et al., 2020). Web Scraping therein includes all methods and techniques for extracting data and information from the web (Dewi et al., 2019). In contrast, API is a back-end interface built for developers that directly communicates with the database of a service. They are thus often described as “glue” that sticks together different computer systems (Sloan & Quan-Haase, 2017). Many social media platforms make data available through APIs, with which they are governed by their technical and legal regulations. This allows researchers to retrieve, store, and process digital traces left by users of the respective platforms for empirical analysis (Perriam et al., 2020).

One of the main advantages of API-based data collection is the easy access. Although a certain technical and programming expertise may be required, a few lines of code are often sufficient for simple API requests. Besides, there is a variety of open-source tools which can be used without the necessity of prior programming experience. By contrast, the aforementioned access method of web scraping generally involves code-based frameworks that are much more difficult to learn (Lomborg & Bechmann, 2014). Not only is the workflow much more complex and time-consuming than API extraction, but it also differs for every website due the pages’ unique characteristics (which also tend
to change over time). Although web scraping is generally more flexible than using APIs, this type of data collection is usually prohibited by the Terms of Services (ToS) of most social media platforms (Halavais, 2019). However, a recent U.S. court ruling indicates that scraping publicly accessible internet data is legal for academic research (Whittaker, 2022). It should still be treated carefully, as it may contain sensitive information and cause the platform to block the IP that is used to scrape data.

Despite its advantages, API-based data collection also brings about several downsides that researchers should be aware of when considering using them. While APIs make data publicly available, they are not open in the sense of giving full and unlimited access to the entire database. APIs are limited to public data¹ and usually also restricted in volume and variety. In most cases, only samples of the data can be obtained with the process behind the sampling usually being unknown or at least not fully transparent to researchers (Morstatter et al., 2013).

In addition, the ToS of platforms and their API are subject to change and can, therefore, be becoming unreliable for continuous (or repeated) data access. Relying on APIs for acquiring data, hence, means that research is, to a large extent, dependent on decisions made by commercial companies (Bruns, 2019; Freelon, 2018). As a result of debates about data security and user privacy following the Cambridge Analytica scandal, many platforms, notably Facebook and Instagram, have continuously restricted access to their data. This has led to a general instability of digital data sources (Gerlitz, 2016; Perriam et al., 2020). The risks and implications that come with what some have called a "Post-API age" (Freelon, 2018) or an "APICalypse" (Bruns, 2019) are severe. API shutdowns or restrictions create a significant gap in data access among researchers and can also hamper the efforts of third-party developers who create tools for collecting and processing these data (Bruns, 2019; Freelon, 2018). Short-term changes or shutdowns of API functionalities also lead to a fragile data access structure that can be too unstable for longitudinal studies (Perriam et al., 2020).

Besides taking into account the risks associated with the use of APIs, before collecting any data, researchers should determine what data are needed for the respective research objective and choose an appropriate platform, sampling approach, and collection tool(s). Regarding the sampling procedure, the most common approaches are user-related, keyword-related (e.g., tracking a specific hashtag) or URL-related (Mayr & Weller, 2017; Stieglitz et al., 2018). Notably, due to technical, legal, or ethical limitations, however, it may be possible that not all available data can actually be acquired. There are several methods for programmatically collecting data from APIs or endpoints. These include commercial services and tools, as well as options that involve programming languages like Python or R, API tools, or free command line tools like curl. Some of these methods may have limitations in terms of transparency or ease of modification.

To facilitate making informed choices for data access, we provide a commented list of tools that can extract social media in the following. As web scraping is harder to learn and usually prohibited by the platforms’ ToS, the focus of the lists is on tools that make use of platform APIs. Since all data collection tools come with certain limitations, choosing the "right" tool strongly depends on the research objective and the researchers’ technical capabilities. The aim of this document is to provide an overview and comparison of the tools. Please note that questions of research ethics and legal perspectives on working with data from social media platforms, such as data privacy, informed consent and data ownership, are not considered in this document.

Additional Resources

General resources


API resources


¹ The definition of „public data“ varies between platforms.
Review: https://bookdown.org/paul/apis_for_social_scientists/ (last access: 17/01/2022)


Web scraping resources


Disclaimer

The team Research Data & Methods at CAIS is not responsible for the tools listed in this document and their operability. In addition, we cannot guarantee that the tools are in compliance with the ToS of the respective platforms. Due to frequent and unforeseeable changes in the APIs, some tools listed here may cease to function. Likewise, tools may not be maintained and actively developed anymore. The list will periodically be checked and updated. Feel free to contact rdm@cais-research.de for further information or if you have suggestions for changes.

Acknowledgement

The majority of the tools listed in this document were identified via the following resources:


Explanation of the table columns

- **Name**: Name of the tool including link
- **API-based GUI**: Specifies whether the tool offers a graphical user interface (GUI)
- **Programming required**: Indicates whether the tool uses APIs of the platform or relies on webscraping
- **Costs**: States whether the users need coding skills and if yes in what programming language
- **Analysis Features**: Some tools offer functionalities for analyzing data. However, these are usually limited in comparison to full-fledged analysis solution.
- **Free (to use) and open source**: FOS – Free to use and open source
- **Free to use, not open source**: Pay – Not open source and with costs
- **Last Access**: Shows when the tool availability was last checked by the authors of this guideline
Navigating Research

Data and Methods

November, 2023

Twitter/X

Over the last years, Twitter has become the “model organism” (Tufekci, 2014, p. 2) of social media research, mainly due to the fact that the data are relatively easy to obtain through the platform’s API. One thing to note here is that a developer account and an API Key (that can be obtained with a developer account) is required for using use the API. The signup to the Twitter Developer Portal can be found here. The API has numerous endpoints that can be used to access various types of data, such as posts, user profiles, followers, or mentions.

In the beginning of 2021, Twitter had launched an Academic Research product track for their new API version (V2) which was available for non-commercial academic research. After applying for and receiving authorization, researchers could gain access to the full history of public tweets and a higher monthly cap. In 2023, significant modifications have been implemented to the Academic Research API, rendering it highly restrictive. Twitter made an announcement in April 2023, stating that accessing the API would necessitate a paid subscription for being able to collect data. Access levels are divided into Free, Basic (100S/month), Pro (5000S/month) and Enterprise. In July 2023, Twitter’s CEO Elon Musk renamed the company from Twitter to X.

API rate limits: The API v2 has a consumption cap which limits the number of tweets that can be received on a monthly basis. The rate limits also differ between the newly introduced access levels. As such, the Basic version has a post cap of 10,000, while the Pro version has a cap of 1,000,000. The free access level cannot pull posts.

Requests are made per 15-minute window, whereas every specific lookup has a different limitation. The requests overview can be found here. Some requests may have a monthly cap that do not fall under the tweet consumption cap described above. For example, the “get users” request has a monthly cap of 15,000 independent of any other rate limits. Please note that packages and tools are constantly being archived due to the fact that open-source solutions are not maintainable anymore due to high API costs by X.

Additional Resources

<table>
<thead>
<tr>
<th>Name</th>
<th>API-based</th>
<th>GUI</th>
<th>Programming required</th>
<th>Costs</th>
<th>Analysis Features</th>
<th>Short Description</th>
<th>Last Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audiense</td>
<td>Yes (V1.1/V2)</td>
<td>Yes</td>
<td>No</td>
<td>Pay</td>
<td>Yes (Consumer Insights)</td>
<td>Utilize social data to understand audiences (marketing focused)</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Chorus Analytics</td>
<td>Yes (V1.1/V2)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>Yes (Visual Analytics)</td>
<td>Free, evolving, data harvesting and visual analytics suite to facilitate and enable social science research</td>
<td>12/09/23</td>
</tr>
<tr>
<td>DMI-TCAT</td>
<td>Yes (V1.1/V2)</td>
<td>No</td>
<td>Yes (PHP)</td>
<td>FOS</td>
<td>Yes (Keyword Analysis)</td>
<td>Robust and reproducible data capture and analysis, and interlinks with existing analytical software</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Facepager</td>
<td>Yes (V1.1/V2)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>No</td>
<td>Fetches publicly available data from YouTube, and Twitter based on APIs</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Python-twitter</td>
<td>Yes (V1.1/V2)</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>No</td>
<td>A Python wrapper around the Twitter API</td>
<td>12/09/23</td>
</tr>
<tr>
<td>RTwitterV2</td>
<td>Yes (V2)</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>No</td>
<td>Code to loop through timelines and the academic search API</td>
<td>12/09/23</td>
</tr>
<tr>
<td>SMO-TMAS</td>
<td>Yes (V1.1/V2)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>Yes (Text &amp; Network Analysis)</td>
<td>Shiny app that makes it easy to collect and analyze small Twitter data sets (&lt;18,000 tweets)</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Scweet</td>
<td>No (Scraper)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>No</td>
<td>Simple and unlimited Twitter scraper using Python</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Social Bearing</td>
<td>Yes (V1.1/V2)</td>
<td>Yes</td>
<td>No</td>
<td>Pay</td>
<td>Yes (Consumer Insights)</td>
<td>Insights &amp; analytics for tweets &amp; timelines</td>
<td>12/09/23</td>
</tr>
<tr>
<td>TAGS</td>
<td>Yes (V1.1)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>No</td>
<td>Google Sheet template which lets you setup and run automated collection of search results</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twacaptic</td>
<td>Yes (V2)</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>No</td>
<td>API client wrapper that automates common tasks (e.g., get all tweets by a list of users and poll for new tweets regularly)</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twarc</td>
<td>Yes (V1.1)</td>
<td>No</td>
<td>Yes (Python/Bash)</td>
<td>FOS</td>
<td>No</td>
<td>Command line tool and Python library for archiving Twitter JSON data</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twarc2</td>
<td>Yes (V1.1/V2)</td>
<td>No</td>
<td>Yes (Python/Bash)</td>
<td>FOS</td>
<td>No</td>
<td>Collects data at the command line from the Twitter API</td>
<td>12/09/23</td>
</tr>
<tr>
<td>tweepy</td>
<td>Yes (V1.1)</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>No</td>
<td>Easy-to-use Python library for accessing the Twitter API</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twitterAcademic</td>
<td>Yes (V2)</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>No</td>
<td>Use R to retrieve data from Twitter API 2.0 academic research track</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twitter-explorer</td>
<td>Yes (V1.1)</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>Yes (Network Analysis)</td>
<td>The twitter explorer combines collection, transformation, and visualization of Twitter data in an easily accessible interface</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Twitterscraper</td>
<td>No (Scraper)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>No</td>
<td>A simple script to scrape Tweets using Python to retrieve Twitter content without API restrictions</td>
<td>12/09/23</td>
</tr>
<tr>
<td>twittercrawler</td>
<td>Yes (V1.1)</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>No</td>
<td>Provides a way to collect network data through Twitter's standard v1.1 APIs</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Twurl</td>
<td>Yes (V1.1/V2)</td>
<td>No</td>
<td>Yes (Ruby)</td>
<td>FOS</td>
<td>No</td>
<td>OAuth-enabled curl for the Twitter API</td>
<td>12/09/23</td>
</tr>
<tr>
<td>Twython</td>
<td>Yes (V1.1/V2)</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>No</td>
<td>Premier Python library providing an easy (and up to date) access to data</td>
<td>12/09/23</td>
</tr>
<tr>
<td>yasonSML</td>
<td>Yes (V1.1)</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>Yes (Text &amp; Network Analysis)</td>
<td>Provides a suite of tools for collecting and constructing networks from social media data</td>
<td>12/09/23</td>
</tr>
<tr>
<td>yasonTcn</td>
<td>Yes (V2)</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>No</td>
<td>Collect tweets and generate networks for threaded conversations</td>
<td>12/09/23</td>
</tr>
</tbody>
</table>
Although YouTube is the largest online video platform, the data it generates have been used much less frequently in academic research (Thelwall, 2018). The YouTube Data API provides access to data, such as video information, comments, playlists, or channel statistics. In order to use the API, a Google account and a project that has to be created via the developer console are required (Breuer et al., in press; Kaplan & Klein, 2022). As of July 2022, the platform provider has introduced the YouTube Researcher Program, which is an initiative that provides eligible researchers scaled access and insights to YouTube's public data as well as support and technical guidance from YouTube.

API rate limits: The YouTube API has a quota of 10,000 units per day, and each API method is associated with a different quota cost. The quota cost for each request can be assessed using the YouTube Quota Calculator. Through the YouTube Researcher Program, eligible researchers can obtain as much quota as they require for their research. Data access may be restricted depending on your research project by YouTube. The extended quota and program access is only operative for the duration of your research and will be revoked after completion (YouTube, 2022). This also comes with support and technical guidance. The application process, however, is a long process for which enough time must be considered before the data collection.

**Additional Resources**


**Facebook**

As a response to the Cambridge Analytica scandal, Facebook essentially closed access to their Pages and Graph API which previously allowed researchers to extract posts, comments and associated metadata from public Facebook pages (Bruns, 2019). Currently, the only possibility for collecting individual-level user data from Facebook is through creating an own application that each user must give permission to access their data. Some types of Facebook data can be accessed through the CrowdTangle API, a public insights tool which was acquired by Facebook in 2016. Apart from Facebook, CrowdTangle also allows users to follow and analyze public content on the social media platforms Twitter, Instagram, and
Reddit. As of 2019, the CrowdTangle API and user interface are also available for academics and researchers after authorization. However, access to CrowdTangle is controlled and Facebook does not allow data from CrowdTangle to be published. While researchers may publish aggregate results from their data analysis, the original data cannot be shared publicly.

**API rate limits:** By default, the CrowdTangle API has a limitation of 6 calls per minute for all data except for links, for which the rate limit is 2 calls per minute. The rate limits of the Facebook APIs amount to 200 calls per hour.

### Additional Resources


### Instagram

Just as the Facebook API, Instagram's public API has been essentially shut down for external access by academic researchers in the wake of the Cambridge Analytica scandal (Bruns, 2019; Meier-Barthold, 2022). The Instagram Graph API, hence, has similar restrictions as the Facebook Graph API. There are two main APIs for Instagram: the Instagram Basic Display API and the Instagram Graph API. Both offer functionalities for collecting public data that may be of interest for researchers.

**API rate limits:** The Instagram Graph API and Instagram Basic Display API have a rate limiting of 200 calls per hour. Prior to the Cambridge Analytica scandal, up to 5,000 calls per hour were possible.

### Further Resources


Reddit

Reddit has become one of the most prominent social platforms on the web and a popular environment for discussing prominent and controversial events (Proferes et al., 2021). Its subreddit structure makes finding relevant data relatively easy, presenting an additional advantage compared, e.g., to Twitter (Amaya et al., 2021; Proferes et al., 2021). In the past, Reddits’ open and free API contributed towards being a popular and valuable data source for researchers, similar to Twitter.

Starting July 1, 2023, Reddit has started to restrict free API access by introducing reduced rate limits. There now exists an Enterprise Level Tier that costs $0.24 per 1K API calls. Upon request, researchers may be provided with higher rate limits.

**API rate limits:** The free Data Reddit API has a rate limit of 100 queries per minute, and 100 items can be retrieved per request.

**Additional Resources**


Other platforms

In the previous sections, we focused on the social media platforms with the largest global user base: Twitter, YouTube, Facebook, and Instagram. There are, however, also tools available for other platforms, such as Telegram, Wikipedia or TikTok. While we cannot cover all of those, we list some examples in the following.
**Additional Resources**


---

<table>
<thead>
<tr>
<th>Name</th>
<th>Platform</th>
<th>API-based</th>
<th>GUI</th>
<th>Programming required</th>
<th>Costs</th>
<th>Analysis Features</th>
<th>Short Description</th>
<th>Last Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaper</td>
<td>Reddit, Tumblr, Pinterest</td>
<td>No (Scrape!)</td>
<td>Yes</td>
<td>No</td>
<td>FOS</td>
<td>Yes</td>
<td>An open source and collaborative framework for extracting the data you need from websites</td>
<td>25/06/23</td>
</tr>
<tr>
<td>PykTok</td>
<td>TikTok</td>
<td>No</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>No</td>
<td>A simple module to collect video, text, and metadata from TikTok</td>
<td>25/06/23</td>
</tr>
<tr>
<td>scholar network</td>
<td>Google Scholar</td>
<td>No</td>
<td>No</td>
<td>Yes (R)</td>
<td>FOS</td>
<td>Yes (Network Visualisation)</td>
<td>Extract and Visualize Google Scholar Collaboration Networks</td>
<td>25/06/23</td>
</tr>
<tr>
<td>Telegram Analytics</td>
<td>Telegram</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>Yes</td>
<td>Index of Telegram channels (not all countries/languages included)</td>
<td>25/06/23</td>
</tr>
<tr>
<td>TikTok-API</td>
<td>TikTok</td>
<td>Yes</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>Yes</td>
<td>Python Wrapper for the TikTok API</td>
<td>25/06/23</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Wikipedia</td>
<td>Yes</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>Yes</td>
<td>Python library that facilitates accessing and parsing data from Wikipedia</td>
<td>25/06/23</td>
</tr>
<tr>
<td>Wikipedia-API</td>
<td>Wikipedia</td>
<td>Yes</td>
<td>No</td>
<td>Yes (Python)</td>
<td>FOS</td>
<td>Yes</td>
<td>Python Wrapper that supports extracting texts, sections, links, categories, translations, etc.</td>
<td>25/06/23</td>
</tr>
</tbody>
</table>
References


Research Methodology, 21(3), 303–316.  
https://doi.org/10.1080/13645579.2017.1381821


https://research.youtube/policies/terms/

How to cite