

## Data Access Resource

### Methodological note:

- To determine data availability, we reviewed the codebooks for all data access modalities provided by the platforms. If a data point was available through any of these modalities, we marked it as available.
- In some cases, a data point was technically available but not practically usable, for example, when researchers could only access aggregated information rather than download the underlying data. These fields are marked with a caution symbol.
- For LinkedIn, we were unable to locate a codebook for the platform's API (Researcher Beta Programme). Consequently, all fields for this modality were marked with a caution symbol.
- As our analysis is based solely on the codebooks, we cannot assess the actual quality or completeness of the data in practice. Evaluating data quality falls outside the scope of this study.

## Glossary

### Social Media Entity

**Account:** A user-specific profile created on a social media platform that allows individuals or organizations to access, participate in, and manage content and interactions.

**Channel:** used to distribute content, often in a broadcast style, to an audience that follows or subscribes.

**Page:** A feature meant to represent and promote an identity (like a brand or person) and engagement with an audience. Often allows two-way interaction (comments, likes, messages from audience).

**Public Group:** An online space where multiple users can share posts and interact with each other. Membership and content are openly accessible. Anyone can join without approval.

**Subreddit:** A specific community or forum within the website Reddit, dedicated to a particular topic, interest, or theme. Each subreddit is identified by "r/" followed by its name (for example, r/science or r/movies) and is managed by its own moderators.

**Supergroups:** An expanded version of a regular Telegram group that supports a much larger number of members (up to 200,000). Supergroups offer advanced management tools, such as message moderation, pinned messages, and improved permissions.

## Types of data

**User data:** information collected about social media users. This includes demographic details (such as age, gender, and location), profile information (like interests and bios), account characteristics (e.g., account age, settings, and usage patterns), and other identifiable or descriptive attributes.

**Content data:** information that users (create and) share on social media platforms, including text posts, comments, images, videos, stories, and any other media shared by users.

**Interaction data:** Information that captures how users engage with content and with each other on a social media platform, for example, actions such as likes, shares, comments, reactions, retweets, mentions, follows, and clicks.

**Metadata:** Data that describes or provides context about content, user activity, or interactions on a platform, without including the content itself. This may include information such as timestamps, geolocation, device type, language, file size, and platform-specific tags, which help structure, organize, and interpret platform activity.

## Data fields

### User data

**ID (Identifier):** A unique alphanumeric code or string assigned to an account, channel, public group, page or other entity on a social media platform. This identifier is assigned and used internally by the platform to distinguish and manage entities.

**Name (Entity):** A unique identifier selected by a user to represent their account, channel, public group, or page on a platform. It is used for logging in and system identification and is publicly visible as part of the social media entity's profile.

**Handle:** A publicly visible name chosen by a user, typically preceded by a symbol such as "@", used to tag, mention, or identify the user within conversations, posts, and interactions on social media.

**Bio/Description:** A short text section on a social media profile, page, channel, or group that provides a summary or overview of the individual, organization, or community.

**Followers Count:** The total aggregated number of users who follow, subscribe to, or join an account, page, channel, or public group on a social media platform.

**Followers List:** A compiled list of users who follow, subscribe to, or have joined an account, page, channel, or group on a social media platform. The visibility of a followers list varies by platform and account privacy settings. Often the followers list is publicly accessible, while others may be restricted or private.

**Following Count:** The total aggregated number of accounts, pages, channels, or groups that a user actively follows or subscribes to on a social media platform.

**Following List:** A compiled list of accounts, pages, channels, or groups that a user actively follows or subscribes to on a social media platform. The visibility of a following list can vary depending on platform settings and user privacy preferences.

**Creation time (Entity):** The date and time when an entity -an account, page, channel, or group- was originally created or registered on a social media platform. While not always visible through the public user interface, this information may be accessible through web scraping.

**Geolocalisation (inferred):** An approximate geographic location of a user, determined through indirect data such as IP address, stated location in the user's profile, or regional activity patterns. Unlike precise, real-time geotagging, inferred geolocation provides a general sense of where a user is located, typically at the city, region, or country level.

**Verification status:** An indicator showing whether a social media account, page, or channel has been officially verified by the platform. Verification is typically granted to confirm the authenticity of notable individuals, brands, organizations, or entities, and is often displayed with a badge.

## Content data

**Original Post/Caption:** Textual content published by a user or entity on a social media platform, used to convey information, context, or commentary. This includes the main text of a post or the accompanying text linked to media such as images or videos (caption).

**Edited Post/Caption:** Textual content on a social media platform that has been modified after its initial publication by a user or entity.

**Post ID:** A unique alphanumeric code or string assigned to a specific post on a social media platform.

**Post Link/URL:** A unique URL that directs to a specific post within the platform’s public interface (e.g., a link that takes you directly to a post on Facebook or TikTok, rather than an internal API endpoint).

**Post creation timestamp:** The exact date and time at which a specific post was originally published on a social media platform.

**Post modified timestamp:** The date and time when a specific post on a social media platform was last edited or updated.

**Downloadable content:** Any media or file associated with a social media post -such as images, videos, audio files, or documents- that can be programmatically accessed and downloaded through the platform’s API.

**Type of content:** The category or format of media included in a social media post, such as video, image, audio, document, or text.

**Post author name:** The visible name of the individual user, page, channel, or group responsible for posting or sharing content on a social media platform.

**Post author handle:** The unique, publicly visible identifier (often preceded by “@”) of the user, page, channel, or group that created or shared a specific post on a social media platform.

**Post revealed geotagging:** The geographic location information explicitly shared or displayed with a social media post, showing where the post was created or is associated with.

**Sponsored or branded post:** A visible marker or tag applied to the post that explicitly indicates the post is sponsored, branded, or paid promotion.

**Generative AI label:** A visible marker or tag applied to social media content that indicates the content was created or significantly assisted by generative AI technologies.

## Interaction data

**Reactions count:** The total number of user-generated reaction responses (such as likes, hearts, loves, or other emotion-based buttons) collected and presented as a single value for a social media post or content.

**Shares count:** The total number of times a social media post or content has been shared, reposted, or forwarded by users.

**Views count:** The total number of times a social media post, video, or other content has been viewed or played by users. Depends on each platform methodology to calculate.

**Comments count:** The total number of comments submitted by users in response to a specific post or piece of content on a social media platform.

**Comments text:** The written content of user-submitted comments responding to a post on a social media platform. This data includes the actual text entered by users and may also contain mentions, hashtags, emojis, or links.

**Saves count:** The total number of times users have saved or bookmarked a post or piece of content on a social media platform for later viewing.

**Quote count:** Number of times a post has been reshared with additional commentary (e.g. quote tweets on X).