# Social Listening Companies and Access to Sensitive Data

Kamya Yadav and Alicia Wanless
Carnegie Endowment for International Peace

May 26, 2022

**Speeding Research on the Information Environment**
**researchaccelerator.org**

## Abstract

We analyzed the data collection and protection practices of 16 social media monitoring (SMM) companies.[1] These companies used APIs, third-party cookie crawlers, and AI-powered systems to collect data from social media platforms, blog posts, forums, news websites, review sites, video sites, and podcasts, among other sources. They collected data on the location, demographic, identification, and content posted by users. Only three of the 16 companies published privacy policies with details on how they protect the gathered data. The two most common data protection methods among the examined companies were technical (such as secure servers and industry-standard encryption) and contractual safeguards. Many social listening and SMM companies disclose few details on their data collection and protection practices. However, based on the social media user data these companies claim to be gathering, it appears that social media platforms have been providing them with extensive access to their data, suggesting that platforms could share similar types of data with researchers. Whether they *should* provide access to the data would depend on the researcher, the research question, and the purpose of research, among other ethical and privacy considerations.

[1] It is important to note that these 16 SMM companies are the most common SMMs, but not representative of all SMM companies. However, these 16 companies provide insight into current data access options offered by platforms and potential data protection mechanisms that IRIE can adopt.

**Funding:** This research was generously supported by Microsoft, Craig Newmark Philanthropies, the John S. and James L. Knight Foundation, and the William and Flora Hewlett Foundation. This work was initiated as part of a joint project between the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace and the Empirical Studies of Conflict Project at Princeton University. **Author contributions:** KY collected the data, drafted the manuscript, and co-wrote the report. AW provided oversight over the research process and co-wrote the report. **Data and materials availability:** All data required to evaluate the conclusions are present in the paper or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

## Executive Summary

1. We gathered data on 16 social listening and social media monitoring companies in order to better understand what data digital platforms were already sharing with third-party service providers. The companies examined included Brandwatch, Dataminr, Meltwater, and Talkwalker, among others.
2. Social listening and social media monitoring (SMM) companies provide tools to help brands improve their marketing and sales through market and consumer research. These tools are widely used and sector agnostic; educational institutions, news organizations, government agencies, corporations, tech companies, and nonprofits use their services.
3. The examined companies used APIs, third-party cookie crawlers,[2] and AI-powered platforms[3] to collect data from social media platforms, blog posts, forums, news websites, review sites, video sites, and podcasts, among other sources.
4. The social media platforms from which they draw data include Dailymotion, Douyin, Facebook, Google, Google +, Instagram, LinkedIn, Pinterest, QQ, Reddit, RenRen, Sina Weibo, Twitch, Twitter, Tumblr, Vimeo, Vkontakte, WeChat, and YouTube. Among these, Twitter, Tumblr, Reddit, and Weibo offered full data firehoses.
5. Data collected were of four types:
   5.1. Content data: Content of public posts, comments, likes, shares, videos, images, hyperlinks, etc.
   5.2. Demographic information: Gender, interests and hobbies, age/date of birth, family status, professional status, educational background, language, etc.
   5.3. Identification data: Information from a user's social media profile (i.e. username, name, profile picture, etc.)
   5.4. Location data: Geolocation of users.
6. Three out of 16 companies published privacy policies that outlined what data were collected from the aforementioned sources, how it was used, and how it was protected.
7. Anonymization or aggregation was not common practice among these companies.
8. They used technical (such as secure servers and industry-standard encryption) and contractual safeguards to protect data.
9. Based on the social media user data social listening and SMM companies claimed to be gathering, it appears that social media platforms were providing these third-party service providers with extensive access to their data, suggesting that they could share similar types of data with researchers.

[2] These refer to third-party companies that collect cookies and index information about users from digital platforms.

[3] For example, Dataminr claims they use their AI platform to collect real-time data;
"Dataminr Pulse: Real-time Alerts for Enterprise Risk Management". *Dataminr* (2022).
https://www.dataminr.com/pulse

## Introduction

The Cambridge Analytica scandal, in which an academic researcher used Facebook data to support targeted political advertisement campaigns, highlighted the challenges of data-sharing and made it more difficult for researchers to access data.[4] Yet platform privacy policies indicate that third-party developers have access to user data for advertising and commercial purposes and that public information on platforms can be indexed by search engines.[5] In this report, we analyze how a sample of companies from one category of third-party service provider–social listening and social media monitoring (SMM) companies that provided tools to corporations, governments, and nonprofits to conduct market and consumer research–accessed data from platforms for commercial purposes.

To do so, we examined 16 social listening and SMM companies between March and April 2022. Table 1 presents the full list of these companies.[6] From each company, we compiled publicly available information on its data sources, types of data collected, and data protection practices. The codebook guiding our data collection and the data set can be accessed in section A.2 of the Appendix to this report.

We found that the included third-party service providers already accessed a vast array of social media user-data. Our analysis suggests that social media companies could provide similar access to researchers studying platforms and the broader information environment. Whether they *should* provide access to the data may depend on the researcher, the research question, and the purpose of research, among other ethical and privacy considerations.

**Table 1: Overview of Social Listening and SMM Companies**

| Company | Social Media Platforms Tracked | Data Collected |
|---|---|---|
| AgoraPulse | YouTube, Twitter, Facebook, Instagram, LinkedIn, Google | Identification data |
| Awario | Instagram, Twitter, Facebook, Vimeo, YouTube, Reddit | Geolocation, content data, identification data |
| Brand24 | Facebook, Twitter, Instagram, YouTube | Identification data, content data |

[4] Granville, Kevin. "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widen". *The New York Times* (19 March 2018). https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html
[5] "About Twitter's API". *Twitter, Inc.* (2022). https://help.twitter.com/en/rules-and-policies/twitter-api; "Privacy Policy". *Tumblr, Inc.* (9 February 2022). https://www.tumblr.com/privacy/en
[6] Our initial list included 19 companies, of which three were not included in the final data set. Cyfe and Zignal Labs were dropped because of the lack of data available on them. BuzzSumo was dropped because it is a product of a parent company included in our data set (Brandwatch).

| | | |
|---|---|---|
| Brandwatch | QQ, Baidu, Twitter, Reddit, and Tumblr | Identification data, content data, demographic data, location data |
| Dataminr | Not available | Identification data, content data, location data |
| Digimind | Google | Identification data, content data, location data |
| Hootsuite | Twitter, Facebook, Instagram, LinkedIn, YouTube, Pinterest | Demographic data, content data, location data |
| Linkfluence | Twitter, Facebook, Instagram, Google+, Sina Weibo, Youtube, Dailymotion, Linkedin, Twitch, ВКонтакте (Vkontakte) | Identification data, demographic data, content data, location data |
| ListenFirst | Facebook, Instagram, TikTok, LinkedIn, Twitter, YouTube, Reddit, Pinterest, Tumblr | Identification data |
| Meltwater | Twitter, Facebook, Instagram, YouTube, Reddit, Twitch, Pinterest, Sina Weibo, WeChat, Douyin | Content data; otherwise unspecified |
| NetBase Quid | Twitter, Reddit, Facebook, Instagram | Identification data, demographic data, content data, location data |
| Sprinklr | Twitter, Facebook, YouTube, LinkedIn, Google+, Instagram, Vkontakte, Sina Weibo, RenRen, QQ | Identification data, demographic data, content data, location data |
| Sprout Social | Twitter, Facebook, Instagram, YouTube, Reddit, Tumblr | Identification data, content data, location data |
| Synthesio | Not available | Identification data, demographic data, content data |
| Talkwalker | Twitter, Weibo. | Identification data, demographic data, content data |
| Zeta Global | Not available | Demographic data, location data |

In the subsequent sections, we discuss the data-collection methodology, offer an overview of the companies listed in our data set, and highlight key insights into their data collection and protection practices. We conclude with a review of our findings and takeaways from the report.

## Methodology

We sourced companies in our data set through Google searches for lists of the most commonly used social listening and SMM tools. Some of the companies included were the subsidiaries of a parent company in our data set (for example, Linkfluence is a Brandwatch subsidiary). Though this list was not exhaustive of all the social listening or SMM tools available, analysis of their offerings provided insights into what data are already being shared by digital platforms.

We collected data on each company through three primary methods:

1. Publicly available information on company websites, including their privacy policies, terms of service, and use cases;
2. Free trials of their services;
3. Contacting customer service representatives online.

All of this was done to determine what types of data third-party service providers had access to and how, if at all, that data was protected.

## Overview of Companies

Companies in our data set offered various products that rely on social media data and other internet sources. These products include:

1. Consumer research: provided insights into consumer opinions
2. Audience analysis: allowed analysis of consumer interests and behavioral data
3. Social media monitoring, dashboard, publishing, and analytics: allowed management and analysis of customers' social media profiles
4. Social listening: provided insights into consumer sentiments and market trends
5. Product marketing, advertising, and distribution: enabled the marketing and sale of customers' products and services

Various actors hired these companies to gather information on consumers and the market. Products offered were marketed for three broad purposes: social listening, social media monitoring, and content and brand management. Social listening refers to services that use social media and other internet sources to gather information on the brand/actor in question, conduct sentiment analysis to understand what users of the brand/actor are talking about, and understand market trends. Social media monitoring allows brands/actors to analyze how consumers are interacting with their content, what type of content is doing well, on which platform, and with what kinds of consumers, among other questions. Content and brand management allows brands/actors to promote themselves and their products and manage multiple social media profiles linked to them. These purposes often interacted with one another in the tools and products offered by the companies in our data set.

These tools were widely used and sector agnostic. They were used by educational institutions, news organizations, government agencies, corporations, tech companies, and nonprofits alike. Table 2 shows a sample of organizations by sector that employed some of the companies listed in our data set, according to service-provider websites.

**Table 2: Sample of Organizations that Use Social Listening and SMM Tools**

| Sector | Organization(s) |
|---|---|
| Education | Stanford University, Georgia State University, West Virginia University |
| Journalism | The Washington Post, The New York Times, CNN, The Economist |
| Government | European Investment Bank |
| Corporations | Microsoft, Google, Spotify, BMW, Audi, Nike, Accenture, Unilever, Netflix |
| Nonprofits | UNESCO, UNICEF |

**Data Collection**

Companies in our data set used APIs, third-party cookie crawlers, and AI-powered platforms to collect data from social media platforms, blog posts, forums, news websites, review sites, video sites, podcasts, email lists, customer relationship management (CRM) data, scanners, and the ‘dark web’, as well as through partnerships with offline data compilers, credit bureaus, and financial institutions. They drew data from social media platforms including Dailymotion, Douyin, Facebook, Google, Google +, Instagram, LinkedIn, Pinterest, QQ, Reddit, RenRen, Sina Weibo, Twitch, Twitter, Tumblr, Vimeo, Vkontakte, WeChat, and YouTube.

Many of these companies offered historical data; the longest range was from Brandwatch, which offered data that went back to 2008. Since some of the companies used platform APIs, they could easily collect historical data on platform users while also utilizing full data firehoses from Twitter, Tumblr, Reddit, and Weibo, among others. Social listening and SMM companies widely stated that they only drew on publicly available social media data. Table 3 illustrates the four types of data collected: location, identification information, content data, and demographic information. Each company may have collected data that belongs within one, many, or all of these categories.

**Table 3: Types of Data Collected by Social Listening and SMM Companies**

| Type of Data | Description |
|---|---|
| Content | Content of public posts, comments, likes, shares, videos, images, hyperlinks, etc. |
| Demographic | Gender, interests and hobbies, age/date of birth, family status, professional status, educational background, language, etc. |
| Identification | Information from a user's social media profile (i.e. username, name, profile picture, etc.) |
| Location | Geolocation of users |

**Data Protection**

Each of the social listening and SMM companies studied distinguished between their customers and the users of digital platforms. All third-party service providers analyzed here published privacy policies pertaining to their customers. However, only a few published privacy policies regarding user data from digital platforms. Out of the 16 companies in our data set, only three had dedicated privacy policies for social media user data: Brandwatch, Talkwalker, and Linkfluence (a Brandwatch company). The website of a fourth, Zeta Global, noted that it only collects data that users opt in to disclosing. The remaining companies did not disclose how they protect the publicly available data collected from digital platforms. This made it challenging to determine how the data are being protected.

From the three available privacy policies covering social media user data, we identified two broad safeguards: contractual and technical. Contractual safeguards consisted of agreements between the companies and their customers placing restrictions on data use and ensuring that the customers' data privacy and protection standards match those of the company. For instance, Brandwatch prohibited "customers from using your Personal Data to target and profile you based on sensitive categories of Personal Data (e.g., health status, sexual orientation, political beliefs, etc.); to single out individuals for unlawful or discriminatory purposes; in any way that goes against the law, including data protection law."[7] Technical safeguards included the use of industry-standard encryption and secure servers to store data.

Most companies in our data set stated that the data were retained for as long as their customers needed it. Several of the companies that had not published social media user data privacy policies stated that the publicly available data they collected was protected by the privacy policies of the digital platforms

[7] "Author privacy statement: 8. How We Protect Your Personal Data". *Brandwatch* (4 April 2020). https://www.brandwatch.com/legal/author-privacy-policy/#how-we-protect-your-personal-data

from which it was gathered and directed customers and users to the privacy policies of those digital platforms. They also emphasized the publicly available nature of data and that they could not access data made private by digital platform users.

**Conclusion**

Social listening and SMM companies provide social media data to companies across sectors. They market their products as tools to conduct trend and sentiment analysis, carry out consumer and market research, and improve marketing and sales. The data in our study were drawn from myriad sources using third-party cookie crawlers and APIs. These data can be very granular; they may include a user's geolocation,  identification information (username, profile picture, name, etc. used on the platform), the content they post, and their demographic information.

Three out of 16 social listening and SMM companies in our data set had privacy policies outlining how they protected social media user data, either through technical (such as industry-standard encryption or secure servers) or contractual safeguards. Though these companies said social media user data were publicly available online, the data they provided is often highly individualized. While demographic and location data may have been aggregated, their customers often had access to individual-level data, such specific content, usernames, and profile pictures of users. Anonymization or aggregation of this data was not common practice among the included social listening and SMM companies.

Though there is not a lot of information available from the examined social listening and SMM companies about their data collection and protection practices, it is clear that they use social media data for commercial purposes. Digital platforms could make that data available to researchers to study platforms and the information environment.

## Appendix

*A.1 HIPAA Data Access*

In addition to social listening and SMM companies, we looked at how private and sensitive health data, protected under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), can be accessed for research. HIPAA ensures the protection of an individual's sensitive health data from being disclosed without the individual's consent or knowledge.[8] The HIPAA Privacy Rule implements the HIPAA provisions and determines the conditions under which protected health information can be used for research purposes. Research can be conducted on data that is either de-identified (the researcher cannot determine an individual's identity from the data) or identifiable (they can). The latter is subject to more conditions and provisions under the Privacy Rule.

There were three categories of data available to researchers: de-identified data; identifiable data obtained with individual authorization; and identifiable data obtained without individual authorization. Access to identifiable data was largely determined either by informed consent and authorization of individuals or by an Institutional Review Board or Privacy Board.

### I. De-identified data

The Privacy Rule stated that, "A covered entity may always use or disclose for research purposes health information which has been de-identified without regard to the provisions below."[9] Covered entities collectively referred to:

1. Health plans
2. Health care clearinghouses
3. Health care providers who conducted certain financial and administrative transactions electronically. These electronic transactions were those for which standards have been adopted by the Secretary under HIPAA, such as electronic billing and fund transfers.[10]

### II. Identifiable data with authorization

[8] "Health Insurance and Portability Accountability Act of 1996 (HIPAA)". *Center for Disease Control and Prevention (CDC) Public Health Professional Gateway* (September 14, 2018). https://www.cdc.gov/phlp/publications/topic/hipaa.html#:~:text=The%20Health%20Insurance%20Portability%20and,the%20patient's%20consent%20or%20knowledge

[9] "Research". *U.S. Department of Health and Human Services Office for Civil Rights (ORC)* (13 June 2018). https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

[10] "Who Must Comply with HIPAA Privacy Standards?". *U.S. Department of Health and Human Services Office for Civil Rights (ORC)* (26 July 2013). https://www.hhs.gov/hipaa/for-professionals/faq/190/who-must-comply-with-hipaa-privacy-standards/index.html

The Privacy Rule allowed covered entities to use or disclose protected health information for research when the individual authorizes the disclosure. The obtained authorization had to satisfy the requirements set forth in the Code of Federal Regulations.[11] Along with a general set of authorization requirements applicable to all uses and disclosures, there were special requirements for research authorizations. For instance, research authorizations may state that the authorization does not have an expiration date, or that the authorization for the use or disclosure of protected health information for a research study may be combined with consent to participate in the research or with any other legal permission related to the study.[12]

## III. Identifiable data without authorization

Under limited circumstances, covered entities could use and disclose protected health information for research purposes without authorization of the individuals. In order to do so, the covered entities must have met one of the criteria listed in Table A1.[13]

**Table A1: Criteria for Use or Disclosure Without Authorization of Individuals**

| Criteria | Details |
|---|---|
| Documented Institutional Review Board (IRB) or Privacy Board Approval | A document that proves the alteration or waiver of research participants' authorization for use or disclosure of information about them for research purposes has been approved by an IRB  or Privacy Board. A covered entity needs to obtain all the documentation outlined under the Privacy Rule before the use or disclosure of protected health information. Documentation includes identification of the IRB or Privacy Board and the date on which the alteration or waiver of authorization was approved, a statement that the IRB or Privacy Board has determined that the alteration or waiver of authorization, in whole or in part, satisfies the three criteria in the Rule, etc.<br>An IRB may approve the waiver of authorization if and only if the following three criteria are fulfilled: |

[11] "45 CFR Subtitle A". *U.S. Department of Health and Human Services* (1 October 2018). https://www.govinfo.gov/content/pkg/CFR-2018-title45-vol1/pdf/CFR-2018-title45-vol1-sec164-508.pdf

[12] A full list of the special requirements for research authorization can be found here under section titled, "How the Rule Works," subsection "Research Use/Disclosure With Individual Authorization": "Research". *U.S. Department of Health and Human Services Office for Civil Rights (ORC)* (13 June 2018). https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

[13] Detailed information for each criteria can be found here under section titled, "How the Rule Works," subsections 1, 2, 3, 4, and 6:  "Research". *U.S. Department of Health and Human Services Office for Civil Rights (ORC)* (13 June 2018). https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

| | |
|---|---|
| | 1. The use or disclosure of protected health information involves no more than a minimal risk to the privacy of individuals.<br>2. The research could not practicably be conducted without the waiver or alteration.<br>3. The research could not practicably be conducted without access to and use of the protected health information. |
| Preparatory to Research | This is used for preparation of a study or research design.<br><br>Representations from the researcher, either in writing or orally, that the use or disclosure of the protected health information is solely to prepare a research protocol or for similar purposes preparatory to research, that the researcher will not remove any protected health information from the covered entity, and that protected health information for which access is sought is necessary for the research purpose. |
| Research on Protected Health Information of Decedents | Representations from the researcher, either in writing or orally, that the use or disclosure being sought is solely for research on the protected health information of decedents, that the protected health information being sought is necessary for the research, and, at the request of the covered entity, documentation of the death of the individuals about whom information is being sought. |
| Limited Data Sets with a Data Use Agreement | A data use agreement entered into by both the covered entity and the researcher, pursuant to which the covered entity may disclose a limited data set to the researcher for research, public health, or health care operations. A limited data set excludes specified direct identifiers of the individual or of relatives, employers, or household members of the individual. The data use agreement must cover permitted uses and disclosures, limit who can use or receive data, and require the recipient of the data to agree to certain criteria. |
| Accounting for Research Disclosures | Individuals whose protected health information has been used or disclosed for research purposes have the right to receive an account of disclosures. |

### *A.2 Codebook*

| Variable | Description |
|---|---|
| Name | Name of the company |

| Variable | Description |
|---|---|
| Website | Company website |
| Products | Company products and tools |
| Data sources (if available) | Where does the company draw data from? |
| Social media companies tracked (if available) | What social media companies does the company draw data from? |
| Data collected | What data does the company collect? |
| Do they disclose data protection methods? | Does the company disclose whether it has policies in place to protect data that they are gathering from digital platforms? |
| Data protection statement | If yes, how do they protect data? (How companies describe their data privacy policies, verbatim) |
| Sample of clients | A sample of their clients (usually available on the website) |
| Free trial available | Does the company offer a free trial for its products? |
| Notes | Any other available information |
| Other relevant links | Any links with relevant information |
| Privacy Policies/Terms of service/Terms of use | Link to Privacy Policy/ToS/ToU when they outline how data are protected, with relevant sections of the document listed |

*A.3 Data*

PE2_Social Listening Companies