Sample DMS Plan: Analysis of social media posts

Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:

1) **Social media posts and comments**: Approximately 300K publicly available social media posts (including images and videos) will be obtained through a third-party social media data mining vendor. The posts will be processed and given a unique identifier; comments will be linked with its associated post/identifier. All identifying information (e.g., names, social media handles or other user identifiers, websites) will be removed. These data will be categorized and coded (e.g. topic, sentiment, accuracy, use of personal narrative).
2) **Survey data:** The study will collect quantitative survey data from 500 young adult participants (age 18-34) enrolled in an intervention delivered via Instagram, a popular social media platform. The survey instrument includes participant demographic characteristics, and measures for all independent and dependent variables. Data will be recoded to characterize missing values and other data processing activities as needed. Each participant will also have their intervention group assignment (e.g., control group, intervention arm) documented. All data will be deidentified.
3) **User analytics:** Program monitoring data such as views, comments, and "shares" from the 500 participants enrolled in the intervention will also be collected with consent. This includes metrics related to each participant's social media engagement within the intervention and control groups on a social media platform, as well as coded results of major themes and characteristics of participants' posts. All identifiers (e.g., names, social media handles or other identifiers, websites) will be removed.

B. Scientific data that will be preserved and shared, and the rationale for doing so:

1) The raw images or texts of original SM posts will not be shared due to third party vendor restrictions and for user privacy protection. The queries and search process, as well as a select number of exemplary posts and codebooks (including variable/data dictionary for each variable) will be shared. Summary data containing post identifiers and associated coded characteristics (e.g. content, sentiment, accuracy), which will be used to complete the aims of the study, will be shared.
2) De-identified survey data and de-identified and coded program monitoring data (e.g., number of logins, "likes", content of posts, other analytics) from intervention participants will be shared.

C. Metadata, other relevant data, and associated documentation:

Documentation to be made publicly available to the research community will include PDF documents containing:

- Search terms, time frame, and other parameters provided to third-party data mining/analytics vendor to collect the corpus of SM posts
- Survey instruments

- Data collection and intervention protocols
- Copies of blank consent forms
- Codebooks for social media data and intervention survey data will be made available as PDFs. Each variable in the codebook will include a brief description of the item along with the question number and question text from the questionnaire, variable name, variable label, value labels, and standard codes for missing values—including codes for non-applicable, “don’t know,” and refusal.

Element 2: Related Tools, Software and/or Code:
All data will be processed and analyzed with STATA and Python and shared in other widely accessible formats, such as SPSS, R, and Excel.

Element 3: Standards:
To facilitate data use, standard processing and documentation protocols adopted by the Inter-university Consortium for Political and Social Research (ICPSR) will be used for data formats and dictionaries as well as for variable names, descriptions, and labels.

Element 4: Data Preservation, Access, and Associated Timelines
A. Repository where scientific data and metadata will be archived:
Public use study data and associated documentation will be made available to the research community through the data repository hosted at ICPSR. ICPSR is a CoreTrustSeal certified repository providing long-term access to, and preservation of data packages curated by domain specialists.

B. How scientific data will be findable and identifiable:
Datasets will be findable and identifiable through a study digital object identifier (DOI) minted by ICPSR. Every ICPSR data collection receives a globally unique and persistent identifier, which are registered with DataCite (a global DOI provider) and included in the citation and metadata record of each ICPSR data collection. ICPSR creates rich study- and variable-level metadata records in the Data Documentation Initiative (DDI) disciplinary metadata format using information supplied by data depositors and other sources. Metadata available for bulk export in a variety of metadata formats (Dublin Core, DDI, and MARCXML), as well as exportable from dataset landing pages, including structured Schema.org data markup indexed by leading search engines. Metadata are organized using standardized, established formats, templates, and vocabularies, and are released with a clear and accessible data usage license.

C. When and how long the scientific data will be made available:
Shared scientific data will be made available no later than the time of an associated publication which uses the scientific data or the end of the award period, whichever comes first. ICPSR currently has no process for deleting or retiring datasets; the scientific data will be made accessible for as long as ICPSR preserves the data.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:
Terms of the contract with the third-party vendor and privacy concern restrict the wide dissemination or sharing of raw social media data. No additional limitations apply other than the controls and privacy protections described below.

B. Whether access to scientific data will be controlled:
Study data will be made available as public use data to the research community via Data Sharing for Demographic Research (DSDR). Users of the public use data must register with ICPSR and agree to the Terms of Use, which are designed to protect study participants by limiting data use to scientific research and aggregate reporting, prohibiting attempts to identify individual SM users or RCT participants, and requiring immediate reporting of any disclosure of study participant identity. Data users also agree not to share or redistribute any data downloads.

C. Protections for privacy, rights, and confidentiality of human research participants:
Once the data collection for this study has concluded, all participant identifiers (e.g., names and addresses) will be removed and destroyed. RCT study participants will be asked to consent to data collection and broad sharing of deidentified data with the wider research community for additional research.

The privacy, rights, and confidentiality of human subject participants will be protected through the removal and redaction/deletion of all direct personally identifying information, the careful classification of any potentially identifying data, and based on the security standards of the DSDR/ICPSR virtual data enclave. These steps will prevent reidentification of study participants and users whose posts are captured in the social media corpus.

Element 6: Oversight of Data Management and Sharing:
Monitoring of and compliance with this Data Management and Sharing Plan will be the responsibility of the project’s Principal Investigator. The plan will be implemented and managed by professional staff working under the direction of the PI.