Implementing Data Collection and Analysis Pipelines in the Social Media Archive at ICPSR

Introduction

The Inter-university Consortium for Political and Social Research's (ICPSR's) Social Media Archive (SOMAR) proposes an Implementation project to build streaming data collection and analysis pipelines. Our project addresses objectives 3.2 and 3.3, listed under "Goal 3: Improve the ability of libraries and archives to provide broad access to and use of information and collections." We request $429,327 from IMLS and will provide $429,327 in cost share. Our project will generate datasets and data services that democratize researcher access to social media data and enable research data users to build and deploy machine learning models in a virtual data enclave (VDE). We will produce code and documentation that empower other archives to adapt our tools for their technical infrastructure and unique data challenges.

Project Justification

How to protect individuals represented in data while enabling data reuse, especially at scale and with computational analysis tools, are fundamental questions facing data archives. Archives are wrestling with a deluge of data from sensors, administrative records, social media and other digital communication tools, and scientific research. In this project, SOMAR implements pipelines for archiving streaming data and facilitating their analysis with computational methods. While SOMAR focuses on data from social media, the collection and analysis tools we will develop will be useful for other types of streaming, large-scale, sensitive data.

Social media data are incredible resources for science. For instance, they have been used to study political communication, public health messaging, social support, and to predict labor market changes. Researchers face myriad challenges when working with social media data – some because of the people and platforms involved in its creation and retention and some because of its scale. Social media data are generated by individuals and groups and reach audiences through platforms such as Facebook, X (formerly known as Twitter), and LinkedIn. Much of the data individuals generate by using social media is technically or legally "public" (meaning that anyone online can see it). However, individual users and regulatory jurisdictions recognize that making something public does not imply that it's acceptable to use the data for research, especially when the data is sensitive or connected to an identifiable individual. For instance, the European Union's General Data Protection Regulation specifically addresses the tension between research and privacy by outlining acceptable research exemptions and practices (e.g., including safeguards, balancing risks and rewards). The United States has no similar privacy law at the federal level, but the public conversation about social media wrestles with this tension between individual privacy and social benefits of data analysis.

Social media platforms have various rules, often called "terms of service" (TOS), that govern access to their data. Many TOS prohibit data sharing, requiring that individuals collect their own datasets. Requiring users to generate their own datasets restricts participation in science that depends on social media data – only researchers with computational skills and resources are able to collect and manage data. The sharing prohibition also limits replicability and validity checks. Common data resources that ensure broad, perpetual, and consistent data access are necessary for science and related activities that depend on social media. Social media data are also large in scale, especially compared to traditional sources of social science data such as surveys. Datasets such as a day of tweets or a year of Reddit posts are too large, both in disk size and in the number of observations, to analyze manually. Computational and statistical tools such as topic modeling, component analysis, and unsupervised classification help researchers make sense of large datasets. However, machine learning models can also present privacy challenges because they can store and leak data used to train them. Running advanced models also requires significant computing resources such as multiple graphic processing units (GPUs) and high-performance computing (HPC) clusters. Computing resources and skills are not equitably distributed among researchers, and their distribution impacts the researchers and research questions that engage with social media data. More equitable, inclusive means for ensuring access to data are essential for broad, diverse data use.

Project Work Plan

SOMAR will develop data collection and analysis pipelines to address the pressing needs listed above. SOMAR's Project Management and Engineering staff will collaborate to produce accessible end-user tools for collection and analysis. We divide our work into two phases:

1. *Active social media data collection*. In Year 1, Project Management and Engineering teams will collaborate to establish an active data collection system, which will access and store publicly visible social media data. We begin with platforms such as Reddit, X, and Stack Overflow, who have recently changed their TOS to make individual researcher data access more difficult. Our collection tools will adhere to web archiving standards as they collect data and metadata. Project Management will curate metadata and dataset files, making datasets discoverable and accessible to researchers through SOMAR's website.
2. *Social media data analysis*. In Year 2, Project Management and Engineering teams will focus on building systems for data enhancements within SOMAR's secure virtual data enclave (VDE). Enhancements include data linkage, word embedding, auto indexing, and novel measure generation, particularly the ability for researchers to apply existing machine learning models to data in the archive. Engineers will also configure OpenSearch and a related end-user graphical user interface within the secure data enclave. The OpenSearch system will provide users a no-code mechanism for searching across datasets to create bespoke samples that include observations from multiple datasets and platforms. These tools enable researchers to analyze data at scale, study phenomena across social media platforms, and conduct longitudinal and comparative analyses that are not currently possible.

SOMAR will provide access to the resources through its existing cloud-based VDE. Instead of creating local copies of data, researchers log into the enclave remotely and conduct their analysis on SOMAR's computing resources. SOMAR's enclave approach helps prevent data leakage and ensures each project can access the right computing resources (e.g., GPUs, right-sized storage). SOMAR creates individualized VDE instances that contain only the minimum data required for the approved research and the right computing and storage resources for their analysis. In addition, SOMAR protects the privacy of individuals represented in the data by conducting manual disclosure risk reviews of all results before researchers may export them from the enclave. After the Implementation project ends, SOMAR will continue data collection, data enhancement development, archive maintenance, and data user support. As a project of ICPSR, SOMAR's ongoing maintenance is supported by a consortium of over 800 research institutions with over 60 years of data archiving and services experience. IMLS recognized ICPSR with the National Medal for Museum and Library Service.

Over the two years on the Implementation project, Principal Investigator (PI) Libby Hemphill will commit ~1% FTE of her time to provide project leadership and direction to SOMAR staff. At 50% FTE, Project Management staff will plan, monitor, and communicate the project deliverables. They will also prepare reports and coordinate with the Engineering staff to confirm plans of action and milestones, including quality assurance tests to verify the completion of each product deliverable. Engineering staff will commit 100% FTE to build, implement, test, and maintain the technical developments. Project Management and Engineering staff will also create documentation for researchers to aid them in successfully navigating the secure data enclave.

Diversity Plan

Dismantling barriers to data access and analysis are the primary goals of this proposal. Shared data resources play a pivotal role in expanding participation in research. SOMAR levels the playing field to ensure that individuals and institutions can access and analyze data from social media regardless of computational proficiency and local resources. The PI is a queer cis woman, and her team includes individuals with disabilities and members of minoritized ethnic groups. The primary goals of the proposal and composition of the team underscore SOMAR's dedication to diversity and underrepresented voices.

Project Results

This Implementation project addresses researchers' demands for large-scale, automated analysis of born-digital data without compromising the privacy or confidentiality of the individuals represented in the data. It will produce two primary results: shared datasets and data enhancement tools. Both will be provided through SOMAR's existing VDE.

Budget Summary

The estimated budget is $858,654, and we are requesting $429,327 from IMLS. This amount of funding includes salaries and wages ($364,031); fringe benefits ($109,209); computer/servers ($77,179), and indirect cost ($308,235). As required, we included $429,327 in cost-share for this proposal.