Phil Castillo

LIS 545B

Term Project - Final Project Final Submission

14 March 2023

**Data and Metadata Profile**

The dataset of “Suicidality on Reddit” is an organization of anonymous Reddit posts depicting some sort of experience relating to suicide and suicidal ideation. The posts are labeled using the Columbia Suicide Severity Rating Scale (C-SSRS) in order to identify the level of severity of the post. The data itself originates from a Reddit thread consisting of 500 posts submitted by anonymous Reddit users. Since the stated intended use of the dataset is to “enable you to identify and refer those at the highest risk of self-harm or suicide in order to hopefully save lives,” these users are the primary stakeholders for this data. Other stakeholders include the original authors of the dataset and the Conte Center for the Neuroscience of Mental Disorders who created the C-SSRS. Interestingly, the contents of the dataset differ between Kaggle and Zenodo. The dataset found on the Kaggle page contains a singular CSV file while the dataset on the Zenodo page contains a total of seven data files: the CSV file, five XLSXs, and one JSON file. The CSV data file contains an Excel spreadsheet of the 500 anonymous Reddit posts and is organized in three columns: User, Post, and Label. This is essentially what the final dataset looks like. The five XLSX files also contain a spreadsheet of the 500 anonymous posts but are divided in five separate individual “batches.” The final data file contains the raw data of the dataset itself. As far as utilization of the dataset, there are no restrictions but more so a condition. It is stated that if the dataset were to be used for research, that the user must credit the original authors of the dataset. Access to the files, with the exception of the JSON file, does not require any specific software. When downloading the dataset, the user will download a zip file that needs to be unzipped and extracted which will permit the user to view the dataset on an Excel spreadsheet. When trying to access the last file, the user will have to utilize file viewer software.

The metadata consisted of several information as well as some vacant areas. The data comes with information of the collaborator, provenance (the original authors of the data), license, expected update frequency of the data, and an activity overview which includes statistics of the dataset such as the number of views, downloads, downloads per view ratio, and a total of unique contributors. However, this dataset does not have metadata regarding author information, temporal/geospatial coverage, or digital object identifier (DOI). All present metadata is easily accessible and found on the “data card” where all information about the dataset is located. The provided metadata has a basic level of comprehension for the most part. Some aspects of metadata may be unfamiliar to some users, such as “temporal/geospatial coverage” or DOI. But since these specific metadata are absent, it may not be a pressing concern to users. There does not appear to be a specific metadata standard being primarily used. However, users are able to download the metadata in several formats.

In regard to improvements in the current data and metadata, one of the tags used for this dataset, “Clothing and Accessories,” does not seem to fit in with the type of information this dataset is conveying. Omitting it or replacing it with a more relevant tag such as “Reddit” or “self-harm” would not only help users discover this dataset more easily, but also permit them to locate datasets more accurately and access the right information. In terms of enrichment, including some data regarding age, gender, ethnicity, and even class could expand on the data already collected. These specific factors could shed more light onto issues regarding mental health and suicidal ideation as marginality and oppressive experiences due to personal identity can heavily contribute to feelings of depression, anxiety, and other mental disorders. Additionally, this can lead or contribute to other kinds of research that may focus on more narrow aspects of this subject, such as unequal access to mental health resources among marginalized communities. This way other researchers can repurpose this dataset to be more personalized for specific groups. Overall, this would enforce the dataset’s intended purpose of targeted suicide prevention but doing so in a way that is more inclusive and expansive. However, it is understood that this may be difficult to maintain consistency over seeing that the data being collected are solely based on the posts of Reddit users. In this process, personal information may be ambiguous and vague unless the post shares explicit information regarding the user’s identity.

Unfortunately there does not appear to be any publications written based on this dataset that are listed or provided. Instead there are two datasets that contribute to the data and findings of this one. However, upon searching the database name on Google, I was met with a plethora of different scholarly articles that have utilized the data from this set or simply referenced it. It does not seem that there are articles that depict research that have built upon the findings of this particular dataset, but it is disseminating the data and its intended utilization: to help prevent suicidality.

**Repository Profile**

The “Collaborative Psychiatric Epidemiology Surveys 2001-2003” (CPES) repository permits access to information surrounding mental disorders and their relevant treatments. It is centered primarily on adult populations in the United States and intended to be utilized for further research relating to identity-based influences on mental health. This is similar to the dataset “Suicidality on Reddit” which was also created in order to collect data surrounding harmful implications and experiences in the deterioration of mental health with the intention to help create preventative measures. I determined this repository to be a good fit for this dataset not only because of their similarity in focus but also because it was broad enough to house a dataset that did not collect data of a specific group. “Suicidality on Reddit” did not contain data of groups of any particular race, gender, age, etc. Since this repository also did not specify any focus groups with mental disorders but rather centralized mental disorder-related data as a whole, I found that this was synonymous to the dataset.

The organization, Inter-university Consortium for Political and Social Research (ICPSR), to which the repository belongs, permits and encourages deposits of digital data submissions. Members and non-members of institutions are permitted use of ICPSR repositories and anyone is able to contribute data. ICPSR emphasizes that its repository services are open to be utilized by researchers, staff, students, citizen scientists, and the like.

ICPSR appears to encourage any sort of data to be submitted to the repository but requires depositors to include all necessary data and documentation to independently read and interpret the collection, including a study-level description. The only “restrictions'' are in regard to how certain data should be submitted. For example, for quantitative data, ICPSR advises those depositing data files to do so as SAS, SPSS, or Stata files with the expectation that “each variable in the data collection should have a set of exhaustive, mutually-exclusive codes” and that variable and value labels “clearly describe the information or question recorded in that variable” (About ICPSR, 2020). It is noted that datasets in other formats are accepted as well but they do not specify what these other formats are other than ASCII files if the data is accompanied with setup files. Additionally, depositors have the option to deposit data for public-use or have restricted-access.

ICPSR provides the deposit process which can be accessed through a video on the site as well as an informational PDF file describing deposit services and processes in full detail. Additionally, it is noted that depositing data submissions are free of charge with promised benefits of ensuring data meet FAIR principles, help with managing confidential human subject data, assistance with obtaining DOIs and citations, and reviews and improvement of data. In addition to the video and the informational PDF file, ICPSR provides user support through an email if depositors require any further assistance or consultation. This appears to provide opportunity for more specific inquiries regarding particular kinds of data that may not be addressed from the aforementioned resources due to its specificity.

The repository details indicate that the metadata standards included are Data Documentation Initiative (DDI) and Dublin Core with a DCC scheme. However, nowhere in the description of the deposition process does it indicate a requirement for data submissions to be deposited with these metadata standards. The protocol is unclear with the possibility of depositors submitting data not in this structure or according to DDI or Dublin Core specifically.

It is noted that the repository’s data access is available for public downloading through the ICPSR website, but only if the data is specified for public-use. Data in sponsored archives are free and accessible to everyone, but data in the membership archive are available for free for members and $550 per study for others. Membership archive data is able to be freely accessible if the curation’s cost is paid by the depositor or another funding entity. As mentioned, depositors have the option to have their data be restricted-access. If this is the case, then the data can only be accessed by “those who have executed Restricted Data Use Agreement with ICPSR and meet all the criteria by ICPSR” (About ICPSR, 2020).

The repository is accessible using a direct file download. There is an option to download only the documentation but it can also be downloaded as a SAS, SPSS, Stata, ASCII, or delimited file. It is noted specifically for this repository that its public-use data files are accessible to the general public and without a required affiliation with ICPSR. However, it also mentions that the repository contains one or more files that are specially restricted and therefore cannot be directly downloaded.

The repository’s DIP includes a general description of the project with its relevant citations, subject terms, and geographic coverage. Also included are the scope of the project which includes some data collection notes as well as the time period of the data collection (2001-2003) and its methodology.

**Additional Information**

*Recommended data citation*

Guar, Manas, Aribandi, Alambo, Kursuncu, Thirunarayan, Beich, Pathak, & Sheth. (2021). Characterization of Time-variant and Time-invariant Assessment of Suicidality on Reddit using C-SSRS (Version 2 (Version 1: http://zenodo.org/record/2667859#.YCwdTR1OIQI (no post-level annotation))) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4543776

*Considerations for long-term preservation*

Out of all the data files in this dataset, the five XLSX files are more prone to obsolescence and will be the most negatively impactful to the dataset as a whole. Losing one or more of these files will result in the dataset being incomplete with the loss of 100 Reddit posts minimum. Converting these files to more sustainable formats such as CSV would reinforce its preservation, decrease the risk of going obsolete, as well as provide additional back-ups for other datafiles. These files are essentially the building blocks and “meat” of the dataset as a whole, so it is imperative to prioritize protections for these files as well as the others.

In the midst of the seven data files in the repository, the JSON file is the only one that requires special software that permits viewing. This could potentially deter access for some individuals, but this issue can be rectified by including files that do not require special viewing software or by converting the JSON file into a different format.

*Copyright License Statement*

This dataset is peculiar in the sense that its components are found separately in Zenodo and Kaggle. The former displayed the five XLSX files and the JSON files while the latter contained the sole CSV data file. Zenodo lists a Creative Commons Attribution 4.0 license while Kaggle has a Creative Commons Zero Public Domain license. The former license permits the redistribution of the data with the same license which contradicts the Kaggle version of the dataset. If there had to be a decision between the two, it would make more sense to utilize the Zenodo files in favor of the one found on Kaggle. The condition of being permitted to use the data with the same license would not greatly affect its utilization, especially considering that there are multiple files that can be replicated and preserved from Zenodo. The Kaggle file is essentially the same as the JSON datafile from Zenodo, so there wouldn’t be much of a loss there.

*Human subject considerations*

The data depicts personal experiences and has been composed of real text posts on a public Reddit thread. While these posts do not exhibit identifiable data about people that would directly compromise their confidentiality, precaution should still be taken to protect the privacies of individuals tethered to the data when applicable. As of current there has been effort made to anonymize the individuals involved in the dataset, limiting the showcased information to the Reddit username and their respective post. However, additional precautions can be taken such as the consideration of redacting personal identifiable information in the posts themselves such as specific names, locations, etc.