**Sample Application for the UMD Reddit Suicidality Dataset, Version 2**

The UMD Reddit Suicidality Dataset, Version 2 was created by researchers at the University of Maryland and is described at <http://umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html>. It was derived from the [2015 Full Reddit Submission Corpus](https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/) (<https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/>). For more complete descriptions see:

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik, ["Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings"](http://aclweb.org/anthology/W18-0603), *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, Louisiana, June 5, 2018.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych'19), 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, June 6, 2019.

Although Reddit is a site for anonymous discussion and additional steps have been taken to de-identify the data, Reddit data (and particularly data related to mental health) is particularly sensitive. As described on the UMD Reddit Suicidality Dataset, Version 2 web page, therefore, a governance process has been put in place in collaboration with the American Association of Suicidology for sharing of the dataset. This document provides the template for the application and for your convenience it also provides sample responses, as well as sample language that can be used when describing your research with this dataset to your organization’s IRB or equivalent ethical review board.

**Part 1. Use of the Dataset**

a. Please briefly describe your intended use of the dataset.

*Example response:*

*We intend to use the dataset to develop and evaluate algorithms that analyze social media posts in order to predict the author’s level of risk for a suicide attempt.*

b. Please specify the intended duration of the use of the dataset. Please limit yourself to no more than two years, at which point you are welcome to re-apply.

*Example response:*

*My team will use the dataset from DATE through DATE.*

c. Please affirm that you and your team have read Benton, A., Coppersmith, G. and Dredze, M. (2017), “Ethical research protocols for social media health research”, in Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (pp. 94-102), http://www.aclweb.org/anthology/W17-1612, and that in your use of this dataset you are committed to maintaining that paper’s broad ethical principles.[[1]](#footnote-2)

*Example response:*

*All participants on my team will read Benton et al. (2017) and we commit to its ethical principles.*

d. Please commit to referring to the "University of Maryland Reddit Suicidality Dataset, Version 2", in any publications using or discussing this dataset using including appropriate references (you can find BibTeX entries on the dataset web page):

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik, ["Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings"](http://aclweb.org/anthology/W18-0603), *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, Louisiana, June 5, 2018.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych'19), 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, June 6, 2019.

Please also commit to acknowledging the assistance of the American Association of Suicidology in making the dataset available, in any publications.

*Example response:*

*My team and I commit to citing appropriate references connected with this dataset (e.g. Shing et al. 2018, the CLPsych 2019 Shared Task overview, etc.) and to acknowledging the American Association of Suicidology in any publications using this dataset.*

e. Please include a brief data management statement indicating the policy (who and how) for your team’s access to and sharing of the data, including provisions for appropriate protection of the data.

*Example response:*

* *(i) The data and any derivatives will be stored only on password-protected servers at [MY ORGANIZATION] where access will be restricted to me and my direct supervisees using Unix group permissions;*
* *(ii) Any copies of the data or derivatives of it will be accompanied by a clear README.txt file identifying me as the contact person and stating further re-distribution is not to take place without contacting me first.*
* *(iii) I will refer any requests for redistribution of the data (outside my own supervisees) to the providers of the dataset, rather than providing the data myself.*
* *(iv) For anyone with access to the dataset, I will maintain up to date documentation of compliance with IRB or ethical review requirements (e.g. I will add new lab members to an approved protocol when necessary, keep copies of any required training certificates, etc.);*
* *(v) On request I will provide a list of names and email addresses for anyone (e.g. my supervisees) who has had access to the dataset, along with any relevant compliance documentation.*

f. Along with this application I am providing documentation of review by my organization’s Institutional Review Board or equivalent ethical review panel.[[2]](#footnote-3) (See Part 2, below.)

Signature by contact person/supervisor:

Affirmed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (please sign)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (please print name)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (please print email address)

**Part 2. Ethical Review by Your Organization**

The dataset was created at University of Maryland, and research on the dataset at University of Maryland has been approved as exempt from Institutional Review Board (IRB) review by the UMD’s IRB, including for sharing with other researchers, in accordance with 45 CFR 46.101(b)(4), which applies to the study of previously existing data that is publicly available *or* involving subjects who cannot be identified directly or through identifiers linked to them. However, because data of this kind can be particularly sensitive, we are also requiring that researchers with whom we share the data obtain ethical review by their own organizations.

For research conducted outside the U.S. or in industry, where an IRB might not be available, an equivalent letter from a research ethics board, IRB equivalent, ministry of health, etc., or permission from the research site will be evaluated on a case by case basis, with the expectation that proposed research will meet U.S. IRB standards as well as any relevant local laws and cultural sensitivities. (Applicants may want to be aware that the U.S. Department of Health and Human Services provides an annually updated [resource to identify federally assured international research sites](https://www.hhs.gov/ohrp/international/index.html), see https://www.hhs.gov/ohrp/international/index.html.)

Every organization has its own template for IRB review, but in order to make things easier, especially for people who have not had a lot of experience with their IRB, we are including example language here for questions that are typically required. Please feel free to use or adapt it however is appropriate for your specific ethical review board.

Type of review

We seek confirmation that our proposed study using the UMD Reddit Suicidality Dataset is Exempt under 45 CFR 46.101(b)(4), since we will be using previously existing data that is publicly available.

Suggested title of the study

Predicting Suicide Risk using Social Media

Abstract/overview

This study will analyze existing publicly available data from Reddit, a web site that supports anonymous online discussions; for example, Reddit’s r/Depression forum includes online discussion of depression. In our study we will use the dataset to develop and evaluate predictive models of suicide risk.

The dataset we will use was created at the University of Maryland for research identifying features of online postings associated with individuals’ mental state and properties of support-oriented online interactions. UMD researchers’ work on this dataset has been approved as exempt from Institutional Review Board (IRB) review by UMD’s IRB, in accordance with 45 CFR 46.101(b)(4), which applies to the study of previously existing data that is publicly available (or involving subjects who cannot be identified directly or through identifiers linked to them).  Reddit forums are public *and* people posting to Reddit are officially anonymous, in that the site does not ask for or link to personally identifying information. UMD dataset adds an additional layer of anonymization by replacing post authors’ self-chosen Reddit identifiers with unique numeric identifiers. In a further step for purposes of extra caution, despite the data being public and the anonymous nature of the site, UMD has further de-identified the text content of post titles and post bodies.

Subject Selection

1. Recruitment: N/A. No new subjects will be recruited.
2. Eligibility Criteria: N/A. No new subjects will be recruited.
3. Rationale: N/A. No new subjects will be recruited.
4. Enrollment Numbers: N/A. No new subjects will be recruited.
5. Rationale for Enrollment Numbers: N/A. No new subjects will be recruited.

Procedures

We will develop predictive computational models using features of posts on Reddit, in order to develop and evaluate technology that might be useful to assist in identifying and monitoring mental state, particularly the risk of suicidality. In support of this goal, we will analyze archival data already collected online in the University of Maryland Reddit Suicidality Dataset.

Risks

The study involves minimal risk. All data that will be analyzed will be public, pre-existing data, with the authors’ self-chosen Reddit identifiers replaced with unique numeric identifiers and further steps taken to de-identify text content. The data will be stored on secure computers with access limited to the investigator and their supervisees.

Benefits

There are no direct benefits to the people whose data are included in the dataset. Gaining insight into aspects of online discussions may contribute to the creation of more effective interventions. This would increase access to therapeutic referrals and care, potentially providing benefit to many people suffering from clinical and sub-clinical mental disorders.

Confidentiality

All the archival data will be stored on secure computers using standard safeguards such as password-controlled access and group-restricted permissions to limit access appropriately. Reddit forums are public *and* people posting to Reddit are officially anonymous, in that the site does not ask for or link to personally identifying information. The UMD dataset adds an additional layer of anonymization by replacing post authors’ self-chosen Reddit identifiers with unique numeric identifiers. In addition, despite the data being public and the anonymous nature of the site, UMD has further de-identified the text content of post titles and post bodies.

Consent Process

Not applicable – analysis of pre-existing, publicly available data.

HIPAA Compliance

Not applicable

1. We are using Benton et al. (2017) in this application process because it is concise provides a nice overview of relevant issues. However, we strongly recommend that all researchers using this dataset also read, and internally discuss, Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceed- ings of the 2nd ACM Conference on Fairness, Ac-countability, and Transparency (FAT\* ’19)*. [↑](#footnote-ref-2)
2. For multi-site research, only one review is needed, by the principal/primary researcher’s organization. [↑](#footnote-ref-3)