Extended Application for Access to the CLPsych 2025 Shared Task Dataset

Background

The CLPsych 2025 shared task uses an extended subset of the CLPsych 2022 dataset ("Reddit-New"), described in https://aclanthology.org/2022.clpsych-1.16.pdf. 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) can still be thought of as sensitive, therefore an application process has been put in place for sharing of the dataset.

Note that this application covers use of the data after the CLPsych 2025 shared task, for associated non-commercial academic research within a specified scope and duration. Any use beyond the approved period or scope requires a separate agreement.

Application

I am applying for access to the CLPsych 2025 shared task data for the following individuals:

No.	Full Name	Email Address
1	Taya Lin	tayalin2018@gmail.com
2		
3		
4		
5		
6		
7		
8		
9		
10		

For requests to extend data access for non-commercial academic research purposes:

Proposed research completion date: 10/30/2025

Research purpose and scope: To build and study a web tool that predicts emotions in social media posts for academic research only.

My team and I confirm the following:

1. The dataset shall be used solely for the specified non-commercial academic research purposes.

2. All dataset files and derivative models shall be permanently deleted upon the approved research completion date. Neither the dataset nor its derivatives shall be copied to external devices or storage for further use.

3. We confirm that we 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 our use of this dataset we are committed to maintaining that paper’s broad ethical principles.

4. We commit to citing the papers that contributed to this dataset in any publications using or discussing this dataset using including appropriate references:

    Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 Shared Task: Capturing Mental Health Dynamics from Social Media Timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

    Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

    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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

    Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

5. The data and any derivatives will be stored only on password-protected servers at ________________
The data and any derivatives will be stored only on password-protected servers at:
on my personal encrypted and password-protected device, accessible only to me.

[MY ORGANIZATION OR LIST MULTIPLE ORGANIZATIONS] where access will be restricted to me and my team using Unix group permissions or a reasonable equivalent, or on cloud-based computational resources (e.g. AWS S3, Hugging Face Hub, Google Colab) which must be configured for private, team-only access with all public sharing capabilities disabled.

6. Any copies of the data or derivatives of it that we create will be accompanied by a clear README.txt file identifying me as the contact person and stating that further re-distribution is not to take place without contacting me first.

7. We will refer any requests for redistribution of the data (beyond the individuals covered in this application) to the providers of the dataset, rather than providing the data ourselves.

8. We commit to <u>NOT</u> submitting all or part of the data to (1) any closed-source model provider or (2) any platform that potentially stores data, including but not limited to ChatGPT and Claude, except when using HIPAA-compliant instances of such services.

Signature by contact person/team lead:

Affirmed: Taya Lin (please sign)

Taya Lin (please print name)

tayalin2018@gmail.com (please print email address)

02.08.2025 (please print signature date)