Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts

Adam Tsakalidis[1,2], **Jenny Chim**[1], **Iman Munire Bilal**[2,3], **Ayah Zirikly**[5],
Dana Atzil-Slonim[6], **Federico Nanni**[2], **Philip Resnik**[8], **Manas Gaur**[7],
Kaushik Roy[7], **Becky Inkster**[2,4], **Jeff Leintz**[9], **Maria Liakata**[1,2,3]
[1]Queen Mary University of London (UK), [2]The Alan Turing Institute (UK),
[3]University of Warwick (UK), [4]University of Cambridge (UK), [5]Johns Hopkins University (US),
[6]Bar Ilan University (Israel), [7]The Artificial Intelligence Institute (Columbia, US),
[8]University of Maryland (US), [9]NORC at the University of Chicago (US)

Abstract

We provide an overview of the CLPsych 2022 Shared Task, which focusses on the automatic identification of *'Moments of Change'* in longitudinal posts by individuals on social media and its connection with information regarding mental health . This year's task introduced the notion of longitudinal modelling of the text generated by an individual online over time, along with appropriate temporally sensitive evaluation metrics. The Shared Task consisted of two subtasks: (a) the main task of capturing changes in an individual's mood (drastic changes-'Switches'- and gradual changes -'Escalations'- on the basis of textual content shared online; and subsequently (b) the subtask of identifying the suicide risk level of an individual – a continuation of the CLPsych 2019 Shared Task– where participants were encouraged to explore how the identification of changes in mood in task (a) can help with assessing suicidality risk in task (b).

1 Introduction

Increasingly the clinical community are looking for new and better diagnostic measures and tools for monitoring mental health conditions. Over the past decade, there has been a surge in methods at the intersection of NLP and mental health, showing that signals for the diagnosis of certain conditions can be found in language. However, most research tasks have been defined on the basis of classifying individuals (e.g., on the basis of suicide risk (Shing et al., 2018; Zirikly et al., 2019) or on the basis of having a mental health condition or not (Coppersmith et al., 2015)), thus lacking the longitudinal aspect of monitoring an individual's mood and well-being in real-time.

Through this shared task we follow Tsakalidis et al. (2022) to introduce the problem of assessing changes in a person's mood over time on the basis of their linguistic content. For the purpose of the task we focus on posting activity in online social media platforms. In particular, given an individual's posts over a certain period in time, we aim: (a) at capturing those sub-periods during which an individual's mood deviates from their baseline mood – a post-level sequential classification task; (b) leveraging this task to help us assess the suicide risk level of the individual – a user-level classification task (Shing et al., 2018) & a continuation of the 2019 Shared Task (Zirikly et al., 2019). Thus, this year's shared task consists of two subtasks: (A) the main task of identifying mood changes in an individual's online posts over time and (B) assessing the suicide risk level of the invididual, where ideally participants will have been able to establish a connection between tasks A and B. This paper makes the following contributions:

- We introduce tasks A and B and provide a detailed description
- We describe the datasets used for these tasks.
- We provide an overview of the secure data enclave environment used for the shared task.
- We provide an overview of participating team selection, evaluation strategy and discussion of results, paving the way for future approaches.
- We present the limitations of the current set up and provide suggestions for future organisers.

2 Task Definitions

Task A involves capturing 'Moments of Change' (MoC) in posts by individuals on social media over time. In particular, following Tsakalidis et al. (2022), given a sequence of chronologically ordered posts between two dates (*'timeline'*) made by an individual on an online social media platform, we aim to capture the post(s) – or the sequence(s) of posts – in the timeline indicating that the individual's mood has shifted in one of the following ways: *(a) Switch* – the individual's mood shifts

{a.tsakalidis,m.liakata}@qmul.ac.uk

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 184 - 198
July 15, 2022 ©2022 Association for Computational Linguistics

suddenly from positive to negative (or vice versa); and *(b) Escalation* – the individual's mood gradually progresses from neutral/negative (positive) to very negative (positive). Both sudden and gradual changes in individuals' mood over time are important for monitoring mental health conditions (Lutz et al., 2013; Shalom and Aderka, 2020) and constitute one of the dimensions to measure in psychotherapy (Barkham et al., 2021). By definition, this task is temporally sensitive, since the goal is to classify each post in a given timeline as belonging to a Switch (IS), belonging to an Escalation (IE) or not being part of either mood shift (O) – with the majority of the posts expected to be (O).

Task B is a continuation of the work by Shing et al. (2018) and Zirikly et al. (2019). Given the posts of an individual, the aim is to classify their suicide risk into (a) no risk, (b) low, (c) moderate or (d) severe level. Due to the very low number of users of (a) and (b) in our data, we have merged the no/low classes leading to a 3-label user classification task. Participants were encouraged to use insights from Task A in solving Task B.

3 Dataset

Dataset creation for the two tasks (§3.4) involved data collection & data relabelling (§3.1), timeline extraction (§3.2) and annotation (§3.3).

3.1 Data Collection

As our ultimate goal is to find the connection between Moments of Change (MoC) in individuals' longitudinal online data (Task A) and other information regarding the individuals' level of risk (Task B), we wanted to repurpose as much as possible existing mental health datasets (Losada and Crestani, 2016; Losada et al., 2020; Shing et al., 2018; Zirikly et al., 2019) by annotating MoC within them. We also collected a new dataset from Reddit annotated for both MoC and suicidality risk. Our final dataset consists of:

Reddit-UMD. The UMD-Suicidality dataset (Shing et al., 2018; Zirikly et al., 2019) consists of 38K posts by 245 Reddit users who have posted in the *r/SuicideWatch* subreddit (and an equal number of control users who do not feature in our tasks). We have labelled the content generated by these individuals with MoC and relabelled the users' risk level for consistency across datasets.

Reddit-New. We collected a new dataset from Reddit, in two steps: we first collected all public Reddit posts in any mental health-related subreddit (MHS) between 2015-2021 (incl.) and then obtained the posting history for 83K users with at least 10 posts in MHS (for the list of MHS, refer to Appendix A).

	Reddit-UMD	Reddit-New	eRisk++	Total
Timelines	90	139	27	256
Users	77	83	26	186
Posts	2,399	3,089	717	6,205
Duration	∼2 months	∼2 months	(varies)	

Table 1: Dataset overview

eRisk++. We obtained the eRisk dataset (Losada and Crestani, 2016; Losada et al., 2020) upon signing a data use agreement.It contains Reddit posts and comments made by 41 users with and 299 users without self-harm conditions. Inspection of posts by the 299 users showed they were irrelevant for our tasks and so we focussed on the 6,927 posts and comments by the 41 users.[1]

3.2 Timeline Extraction

For each dataset, we extracted user timelines to allow annotation of MoC (Task A), while ensuring that these timelines also contain the information required for Task B (i.e., all associated users' posts in *r/SuicideWatch* are included in the timelines). Table 1 provides an overview of the datasets.

Reddit-UMD. We ordered each user's posts chronologically, identified their posts in *r/SuicideWatch* and defined a user timeline as t days around each such post. Upon experimentation t was set to 30. We extracted 156 timelines of [10,125] posts each, so that annotation was manageable, corresponding to 126 users. These timelines were manually inspected internally by two researchers asked to judge the suitability of the former for Task A. Timelines were thus independently labelled as 'good', 'medium', or 'bad' (Cohen's κ=.66).[2] We only kept 90 timelines that (a) were labelled as 'good' by both annotators and (b) contained all of the user's posts on *r/SuicideWatch* so that we could follow the same annotation for Task B as in Shing et al. (2018).

To inform subsequent data collection we analysed what constitutes a 'good' timeline in Reddit-UMD. For this we trained a Logistic Regression learning to separate between 'good' and 'bad' timelines. We used the timeline-level features

[1] As opposed to Reddit-New and Reddit-UMD, the eRisk dataset contains posts *and* comments made by the users on Reddit. For consistency, we will refer to all of them as 'posts'.

[2] Details of the annotation are provided in Appendix B.

[#posts, % of posts in MHS, and % of posts in r/SuicideWatch, r/depression and r/AskReddit[3]], further accompanied by the average difference (in terms of #posts) between two postings on the same subreddit. We found that the % of posts in MHS is the most predictive feature, with 95% of the 'bad' timelines containing less than 17% MHS posts, whereas 99% of the 'good' timelines have contain less than 82%. We use this information to select 'good' timelines for the Reddit-New dataset.

Reddit-New. Following our notion of 'good' timelines in Reddit-UMD we looked for two-month periods within which the user had at least 10 and no more than 125 posts, at least (most) 17% (82%) of which is posted on a MHS. 150 such timelines were selected at random (from an overall of 1,114) and annotated internally for quality (good/medium/bad), similarly to Reddit-UMD – this time by a single annotator, given the high agreement achieved in Reddit-UMD, resulting into 139 'good' timelines (83 users). Interestingly, one timeline in Reddit-New was identical to another one present in Reddit-UMD – signalling a consistency between the collection process of the two datasets – and hence removed from Reddit-New on our final processing.

eRisk++. Two annotators with experience in mental health research on social media independently reviewed 103 timelines to check suitability for task A. 91 timelines were labeled either as 'good' or 'medium' (Cohen's κ=.78). For consistency with the other datasets, we kept the 15 timelines (14 users) having at least (most) 10 (125) posts.

Upon inspecting the resulting datasets, we found that there was a disproportionate representation of 'low' and 'no' risk users based on the labelling provided in (Shing et al., 2018; Zirikly et al., 2019). To mitigate this, we enriched the eRisk++ dataset with 12 timelines by 12 users from UMD-Suicidality, who had been labelled as 'no'/'low' risk in Zirikly et al. (2019). Though we did not use their associated suicidality risk labels, this step ensured a fairer representation of users for capturing MoC (task A).

3.3 Annotation

Task A. We hired four annotators (2 native English, 2 fluent English language speakers), two of whom had previous experience with performing task A on a different dataset (TalkLife), and provided them with the guidelines from Tsakalidis et al. (2022). Briefly, the task involves reading one timeline at a time in an annotation interface and labelling (a) the first post that signals a 'Switch' (IS) in an individual's mood, along with the respective duration of the Switch (range of consecutive posts), as well as (b) the post signalling the 'peak' (most intense posts) of an 'Escalation' (IE) in an individual's mood, along with the respective range of consecutive posts that belong to the same Escalation. The training of the two non-experienced annotators involved annotating timelines from TalkLife that were previously annotated by the two experienced annotators, measuring their agreement and discussing cases of disagreement in iterative cycles, until reaching an agreement level similar to that in Tsakalidis et al. (2022). Subsequently, the four annotators were provided with 10 separate timelines extracted from UMD Suicidality for training purposes, and disagreements in their annotations were discussed in two meetings. Finally, they were provided with the 255 timelines that have been used in the current Shared Task.

Task B. We worked with four Clinical Psychology experts, all of whom are fluent English language speakers. The experts were provided with the guidelines by Shing et al. (2018), which focus on the task of classifying the suicide risk level (no/low/moderate/severe risk) of an individual, solely on the basis of their *r/SuicideWatch* posts. An annotation interface was developed, where the experts could view and assign a single label to an individual based on up to 5 *r/SuicideWatch* posts made by the individual within the Reddit-New and Reddit-UMD datasets. Our experts re-annotated the suicidality risk of users in Reddit-UMD to provide annotation consistency between the two datasets. [4] For users with more than 5 posts on *r/SuicideWatch*, the annotation was performed in several passes, with the most 'severe' label being finally assigned to the respective individual (Shing et al., 2018). We completed two training rounds with the experts, where they discussed disagreements in their labelling and clarified points especially concerning the distinction between 'moderate' and 'severe' cases.

[3] We selected these 3 subreddits on the basis of being present in at least 20% of the timelines.

[4] We did not use the users from the eRisk++ dataset for Task B: the information on the type of subreddit where a post was shared was not available in eRisk and the remaining 12 timelines from UMD-Suicidality (part of eRisk++) were incorporated at a latter stage.

	Task B: #users					Task A: #posts			
	N/A	Low	Mod.	Sev.	Total	IS	IE	O	Total
Train	22	11	55	61	149	327	773	4,043	5,143
Test	4	3	14	15	36	83	208	762	1,052

Table 2: Summary of the data for both tasks.

	Half	Majority	Perfect
Switch (IS)	.451	.264	.129
Escalation (IE)	.550	.309	.122
None (O)	.920	.832	.692

Table 3: IAA for Task A per agreement threshold.

3.4 Resulting Dataset

Task A. Following Tsakalidis et al. (2022), we assess the inter-annotator agreement (IAA) based on the Intersection over Union for each label independently. The majority agreement (see Table 3) is lower than the agreement in Tsakalidis et al. (2022) (.30/.50/.89 for IS/IE/O, respectively), primarily because in the latter there were 3 annotators employed (requiring 2/3 to agree) whereas here a majority requires agreement between 3/4). A post receives the label assigned to it by the majority. In the case of ties the least populous class receives the label (e.g. if 'IS' ('IE') is chosen over 'O'. In the rare (64 cases overall) of a tie between 'IS' and 'IE', we assigned the label 'IE' given its higher prior.

Task B. The agreement between the expert annotators was considerably lower than that reported in Shing et al. (2018) (Krippendorff's α .43 vs .81), primarily for two reasons: (a) in this dataset, there was only one user assigned 'no risk', which is the easiest category to identify even for non-experts; (b) the experts in Shing et al. (2018) had a background on suicidality whereas our clinical psychologists have broader expertise. Most cases of disagreement involved 'moderate' vs 'severe', or 'low' vs 'moderate' as opposed to 'low' vs 'severe'. We used the majority label for each user and in case of ties the highest level of risk assigned was chosen. We split the data into train and test sets (80/20) preserving the distribution of labels in the two sets. Subsequently, all 204/51 timelines from users in our train/test split, were assigned to the respective set (see Table 2).

4 Working in a Secure Environment

The CLPsych shared task 2021 (Macavaney et al., 2021) was the first to be conducted in a secure environment to provide a high level of safety for sensitive data. We have also opted for carrying out this year's shared task in the same secure environment and continue efforts in protecting highly sensitive data. NORC is an independent non-profit research institution at the University of Chicago who provide the NORC Data Enclave(r), chosen both this year and last for the shared task. Compared to other solutions (see for instance Arenas et al. (2019)) the NORC Data Enclave(r) (hereafter, 'DE') does not rely on dedicated laptops but solely on a browser interface over HTTPS channels and Citrix HDX technology, making the setup of a shared task more feasible. All teams (see §5) signed a data use agreement (DUA) and terms and conditions (T&C) with NORC before being provided with instructions to set up multi-factor authentication for login, procedures for requesting the ingress in the DE of written code, libraries, models or additional data and procedures for technical support. All ingress of information into the DE requires thorough system scans and human review to ensure the safety and integrity of the Enclave.

After login authorized users can access a secure virtual machine within the DE. Although all applications and data run on servers in the NORC data center, the user interface is a familiar full Windows 10 virtual desktop. The DE is a closed environment: it does not have access to the internet and all functionalities for moving data in and out of the virtual space are disabled. This Citrix-based technology is configured to prevent users from downloading output from the remote server to an external machine. Similarly, other security protection features prevent the user from using the "cut and paste" feature in Windows to move data from the Citrix session into an Excel spreadsheet residing on the local computer. In addition, the user is prevented from printing the data on a local computer. There is documentation regarding the virtual environment and how to securely connect to the dedicated DE Cluster on Amazon Web Services (AWS). To connect to the cluster (via ssh) users rely on PuTTY and on the dedicated machine they can find a dedicated Python 3.9.1 environment with all requested libraries available (see §5). Users can both run code and submit batch jobs using the Slurm cluster management while also monitoring the budget available for computational experiments. Following last year's suggestions, we ensured participants would be able to use Jupyter Notebooks to implement code on the cluster through ssh tunneling and by opening the notebook in the browser of the Windows machine. At the end of the Shared Task, each team was to inform NORC to egress the predictions for the test set.

Due to an unprecedented technical issue out of NORC's control, several teams faced issues with running their code a week prior to the system submissions deadlineTo avoid eliminating the teams despite their continuous efforts throughout the Shared Task, we decided to distribute the data outside the DE during the last few days on the basis of the signed DUA. To ensure fairness, we asked all teams (i.e., not only the ones affected) to let us know if they would like to receive the data outside the enclave to help them with the system submission. We made it clear that those submitting their results within the DE would feature separately in our evaluation (see Tables 4-5), since they had more limited resources at their disposal.

5 Call for Participation – Teams Selection

We invited teams to register their interest in the shared task by providing details such as team members, motivation, related background, experience and NLP skills. We also asked for their requirements in terms of programming languages, libraries and pre-trained language models to prepare the set up in the DE. Given our limited resources pertaining to the functional costs of using the DE, we were limited to accepting 15 teams (~50 members) for participating in the Shared Task. Therefore, we compiled a list of criteria that were given to two internal reviewers, along with the (anonymised) registrations of interest. The criteria were related to (a) the relevance of the team's background/current work to the shared task, (b) their motivation and likelihood of committing to the task and (c) details provided wrt technical requirements (see Appendix C for the complete guidelines). Based on the reviewers' assessments, we selected 13/37 teams to participate and asked another five applicant teams to be merged together into two groups, so as to accommodate as many requests as possible (one team was formed by three individual applicants, and another individual applicant was merged into a two-member team), leading to the acceptance of 18/37 requests (53 individuals).

6 Evaluation metrics

Task A. Following Tsakalidis et al. (2022), besides the common post-level evaluation metrics (Precision, Recall, F1) – per class and macro-averaged – we report two sets of timeline-level metrics based on work in change-point detection (van den Burg and Williams, 2020) and image segmentation (Arbelaez et al., 2010), emphasizing respectively performance at the level of a timeline and the prediction of regions of change.

Firstly, working on each timeline and label type independently, we calculate Recall $R_w^{(l)}$ (Precision $P_w^{(l)}$) by counting as "correct" a model prediction for label l if the prediction falls within a window of w posts around a post labelled as l in our ground truth – however, a post's predicted label can only be counted as 'correct' only once (at most). By increasing the value of w, we perform a less strict evaluation of a model. Results are macro-averaged for each label independently across all timelines.

Secondly, we assess model performance on the basis of its ability to capture *regions of change*. For each true region $R_{GS}^{(l)}$ within a timeline, we define its overlap $O(R_{GS}^{(l)}, R_M^{(l)})$ with each predicted region $R_M^{(l)}$ as the intersection over union between the two sets. Finally, we retrieve recall- and precision-based *coverage* metrics (again, macro-averaged across all timelines for each label independently:

$$C_r^{(l)}(M \rightarrow GS) = \frac{1}{\sum_{R_{GS}^{(l)}} |R_{GS}^{(l)}|} \sum_{R_{GS}^{(l)}} |R_{GS}^{(l)}| \cdot max_{R_M^{(l)}} \{O(R_{GS}^{(l)}, R_M^{(l)})\},$$

$$C_p^{(l)}(M \rightarrow GS) = \frac{1}{\sum_{R_M^{(l)}} |R_M^{(l)}|} \sum_{R_M^{(l)}} |R_M^{(l)}| \cdot max_{R_{GS}^{(l)}} \{O(R_{GS}^{(l)}, R_M^{(l)})\}.$$

Ideally we want to see a system performing well on both window based and coverage metrics.

Task B. We use standard classification metrics (Precision, Recall and F1) for each user-based class label and macro-averaged. Due to the low number of users in the 'Low' class on the test set, we also report micro-averaged metrics; however, these are added for completeness purposes in our analysis (i.e., the teams were guided to improve their performance on a per-class and macro-average basis).

7 Shared Task Results

This section outlines the submissions by each team. For Task A, we also provide the results of three baselines: the majority classifier, a logistic regression (LR) trained on tfidf features, and BERT trained using the focal loss on a related but separate dataset on the same task (Tsakalidis et al., 2022). For Task B, we include the majority classifier and a LR trained on tfidf features from users' posts.

7.1 Overview

Task A. Each team was allowed to submit up to three sets of test results. Nine teams submitted their

		DE	Post-level Evaluation												Coverage-based Metrics							
			macro-avg			IS			IE			O			macro-avg		IS		IE		O	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Baseline	**Majority**		–	.333	.280	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.142	–	.000	–	.000	.489	.426
	LR-tfidf		.545	.495	.492	.222	.024	.044	.569	.514	.540	.844	.948	.893	.378	.425	.111	.008	.284	.504	.738	.762
	BERT$_f$-TalkLife		.523	.386	.380	.091	.012	.022	.723	.163	.267	.754	.983	.853	.260	.204	.025	.007	.226	.094	.529	.513
System Submissions	**BLUE**		.505	.495	.499	.175	.171	.173	.484	.433	.457	.855	.882	.868	**.499**	.378	**.500**	.028	.299	.395	**.699**	.712
	IIITH		.520	**.600**	.519	.206	**.524**	.296	.402	**.630**	.491	**.954**	.647	.771	.347	.405	.254	**.356**	.249	.373	.536	.486
	LAMA		.552	.535	.524	.166	**.354**	.226	.609	.389	.475	.882	.861	.871	.376	.441	.253	**.373**	.193	.244	.680	.706
	NLP-UNED	✓	.493	.518	.501	.189	.293	.230	.414	.471	.440	.876	.791	.832	.306	.401	.244	.304	.134	.330	.541	.569
	UArizona	✓	.525	.507	.510	.142	.220	.172	.561	.423	.482	.872	.879	.876	.418	.416	.368	.248	.202	.285	.682	.716
	UoS		**.689**	**.625**	**.649**	**.490**	.305	**.376**	**.697**	**.630**	**.662**	.881	**.940**	**.909**	**.506**	**.503**	**.453**	.343	**.369**	**.450**	.695	**.717**
	uOttawa-AI		.505	.530	.512	.213	.244	.227	.402	.553	.466	**.899**	.793	.842	.348	.453	.272	.317	.176	.417	.595	.625
	WResearch	✓	**.625**	.579	**.598**	**.362**	.256	**.300**	**.646**	.553	**.596**	.868	**.929**	**.897**	.472	**.503**	.406	.318	**.307**	**.467**	**.703**	**.725**
	WWBP-SQT-lite		.508	.509	.508	.231	.220	.225	.440	.462	.451	.852	.845	.848	.336	.376	.270	.224	.186	.321	.551	.583

Table 4: Task A – System evaluation, with **first**, **second** and third highest scores (as well as the highest scores for submissions within the DE) being highlighted. Only the best submission for each team is shown, selected separately on the basis of macro-avg F1 (Post-level Evaluation) and F1=2·C$_p$·C$_r$/(C_p+C_r), macro-based (Coverage-based).

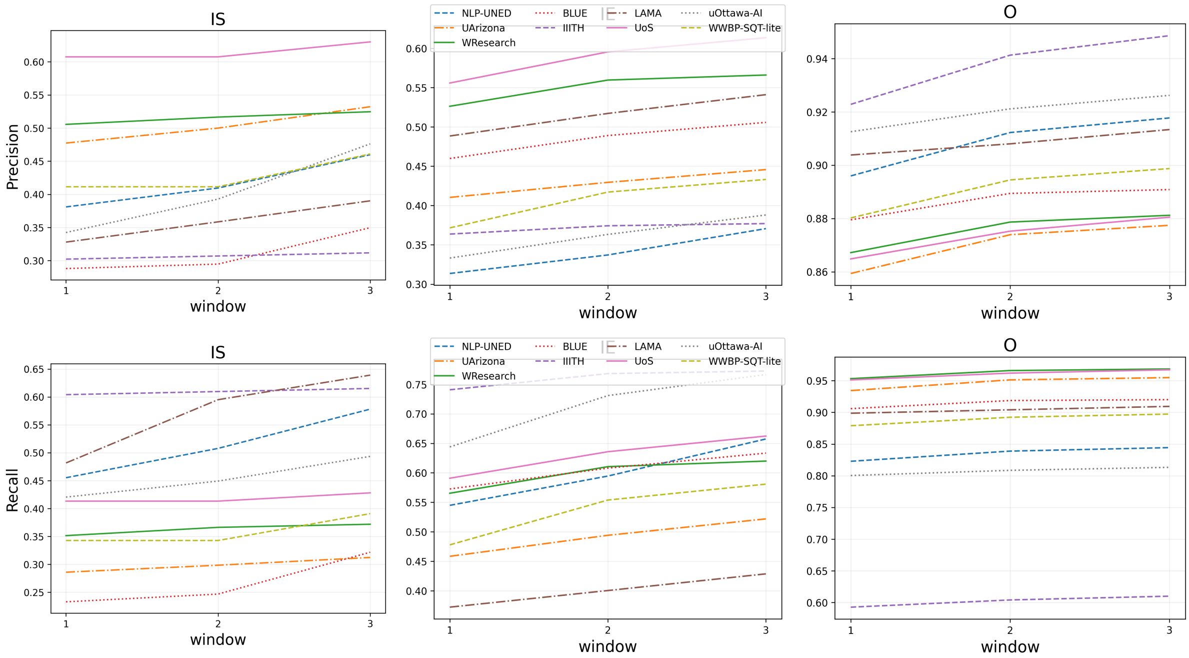


Figure 1: Timeline-level Precision P_w and Recall R_w of the submitted systems. Only the best performing submission by each team is shown (selected on the basis of F1=2·P$_1$·C$_1$/(P_1+R_1), macro-based).

predictions – an overview of the best results per team/metric is shown in Table 4 and Fig. 1. The two best-performing teams (one submitting within and one outside the DE) incorporated a longitudinal component in their models, either in a multi-task setting (UoS) or in an emotionally-informed seq2seq-based approach (WResearch), demonstrating the importance of temporally-sensitive modelling as opposed to classifying each post in isolation. The class imbalance problem was tackled by several teams either via balancing the instances (e.g., LAMA, uOttawa) or via weighted loss functions, notably by IIITH who achieved high recall for IS/IE. Time-related information was incorporated by UArizona, a proximity-based approach was followed by NLP-UNED, an ensemble on emotional and non-emotional features was chosen by BLUE, whereas WWBT-SQT-lite achieved high accuracy (albeit post-deadline) by using different combinations of consecutive post representations.

Task B. Each team was allowed to make a single submission; a second submission was allowed only for teams making use of their predictions from Task A. Seven teams submitted and two teams further took up the challenge of leveraging Task A (see Table 5). The teams that took up this challenge did not demonstrate (important) performance gains. However, the best-performing teams (in average, macro-terms) used some information from Task A, either by focusing mostly on posts labelled as MoC (WResearch) or by jointly learning the two tasks (UoS). The ranking of the teams differs when considering the micro-F1, due to the low number of ‘low’ risk users. Here IIITH and NLP-UNED, along with WResearch, were ranked amongst the top, being particularly effective in capturing ‘severe’ and ‘moderate’ cases, respectively.

7.2 Summary of System Submissions

BLUE (Bucur et al., 2022) explored a variety of feature representation approaches for Task A: (a) Emotion-aware embeddings and (b) non-emotion embeddings (e.g., tfidf, GloVe). They experimented with different combinations of algorithms and features sets, with the most notable performance achieved by a majority voting-based model over an ensemble of predictions obtained by LR, SVM, and Adaptive Boosting classifiers trained on (a), which ranked them second in macro-avg precision-oriented coverage (.499).

IIITH (Boinepelli et al., 2022) used transformers for representing the user’s posts before feeding them to an LSTM for Task A. They tuned their model using the weighted cross-entropy loss function, yielding very high recall for the two minority classes (see post-level results for IS/IE in Table 4). For Task B, they fine-tuned RoBERTa on the training data, tackling the class imbalance with weighted random sampling and producing the outcome label through majority voting. The team came second (third) in this task on micro-F1 (macro-F1), achieving the best scores for the ‘Severe’ class (see Table 5, ‘Severe’).

LAMA (AlHamed et al., 2022) tackled the data imbalance problem by undersampling posts with high sentiment polarity corresponding to the majority class. They adopted a post-level BERT and LSTM models that take into account the sequence of the previous posts for a given target post for Task A. BERT performed particularly well wrt the recall-oriented metrics for IS, leading to the third-best performance in terms of macro-F1 overall. Their models for Task B were Random Forests enriched with sentiment-related features and word frequencies of manually collected high-risk keywords.

NLP-UNED (Fabregat et al., 2022) completed all 5/5 submissions via the DE. In Task A, they analysed the encoded user posts via an Approximate Nearest Neighbour approach – labelling individual posts based on their proximity to others – achieving high recall-oriented scores for IE/IS and the highest macro-average timeline-level recall (for $w = 3$). For Task B, they represented each post on the basis of its proximity to each of the labels in Task A and fed the resulting sequence into a BiLSTM. Amongst the two submissions that leveraged Task A for performing Task B, NLP-UNED was marginally the best-performing in terms of F1.

UArizona (Culnan et al., 2022) completed their 2/2 submissions for Task A via the DE. They tested several variants of RoBERTa-based models, including (a) timeline-agnostic models that incorporate the time lag between consecutive posts and (b) models combining consecutive post vectors, either through concatenation or by passing them through an LSTM to extract the resulting states. They showcased that the incorporation of time boosts the performance of the model on IS cases, whereas they were consistently among the top-3 performing systems in macro-averaged, timeline-level precision.

UoS (Azim et al., 2022) achieved the highest scores for Task A in most metrics and across classes, as well as the second-highest macro-F1 for Task B. They first represent a post in different ways (merged), including its emotion/sentiment-based scores. Their approach involved an attention-based, multi-task BiLSTM operating at the timeline-level, with each post corresponding to a single timestep in the input/output for Task A, and additional outputs for the user’s risk label for Task B at the timeline level (selecting the most ‘severe’ label across all timelines for the user’s classification).

uOttawa-AI (Buddhitha et al., 2022) employed convolutional neural networks with global max-pooling and linear layers for multi-task learning. Task A was casted as two post-level binary tasks (i.e., (a) IS vs O and (b) IE vs O) using soft and hard parameter sharing, by also tackling the class imbalance through down-sampling the majority class. They achieved high recall-oriented metrics for capturing IE and were among the highest scoring teams wrt recall-oriented coverage. In Task B, the team experimented with the additional task of predicting self-declared mental health diagnoses using a separate dataset (Cohan et al., 2018).

WResearch (Bayram and Benhiba, 2022) completed 4/5 submissions in the DE. In Task A, they derived emotionally-informed vectors from pre-trained models and constructed abnormality vectors (i.e., differences in expected vs predicted vectors via a seq2seq model) and differences in the vectors of consecutive posts, using them as inputs to post-level classifiers that take into account the class imbalance. Their best performing submission used

		DE	macro-avg			micro-avg			Low			Moderate			Severe		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
(a)	**Majority**		.156	.333	.213	.220	.469	.299	–	.000	.000	–	.000	.000	.469	1.000	.638
	LR-tfidf		.303	.338	.295	.413	.469	.406	.000	.000	.000	.429	.214	.286	.480	.800	.600
(b)	**IIITH**		.397	.408	.380	**.538**	**.563**	**.520**	.000	.000	.000	**.625**	.357	.455	.565	**.867**	**.684**
	LAMA		.306	.424	.298	.359	.344	.316	.167	**.667**	**.267**	.250	.071	.111	.500	.533	.516
	NLP-UNED (1)	✓	.361	.394	.369	.492	**.531**	.500	.000	.000	.000	.500	**.714**	**.588**	**.583**	.467	.519
	UoS		**.618**	**.427**	**.451**	.482	.469	.438	**1.000**	**.333**	**.500**	.375	.214	.273	.478	**.733**	.579
	uOttawa-AI		.329	.365	.344	.449	.500	.470	.000	.000	.000	.462	.429	.444	.526	.667	.588
	WResearch (1)		**.467**	**.479**	**.465**	**.565**	**.531**	**.543**	**.200**	**.333**	.250	.533	.571	.552	**.667**	.533	.593
	WWBP-SQT-lite		.346	.370	.354	.471	.500	.480	.000	.000	.000	.500	**.643**	**.563**	.538	.467	.500
(c)	**NLP-UNED (2)**	✓	.367	.387	.365	.497	**.531**	.497	.000	.000	.000	**.600**	.429	.500	.500	**.733**	.595
	WResearch (2)	✓	.367	.365	.362	.499	.500	.494	.000	.000	.000	.545	.429	.480	.556	.667	**.606**

Table 5: Task B - System Evaluation: (a) baselines, (b) system submissions, (c) systems utilising Task A.

XGBoost (Chen and Guestrin, 2016) and was consistently among the highest-scoring systems across metrics – and the best-performing from systems within the DE. In Task B, they used LR on n-grams and emotion bandwidth-based vectors extracted from the IS/IE posts for each user, achieving the highest averaged F1. They further leveraged the posts predicted for Task A as IS/IE via a timeline-level BiLSTM, assigning the most 'severe' label for a user based on their timeline classifications, without improvement in performance, however.

WWBP-SQT-lite (Ganesan et al., 2022) experimented with theoretically-motivated features and representations based on Human-aware Recurrent Transformers (Soni et al., 2022) and PCA-reduced RoBERTa. After the deadline the team also tested a version of PCA-reduced RoBERTa vectors, yielding very high accuracy when concatenating them with the previous post's vector and their difference, as features (macro-F1: .61, not reported in Table 4). For Task B the team used LR on user-level features (ngrams, theoretically motivated features), achieving the second-best results on separating the 'Moderate' cases of risk level.

8 Conclusion

We presented the overview of the CLPsych 2022 Shared Task, focusing on (A) capturing changes in an individual's mood as self-disclosed online and (B) classifying the individual's suicide risk level – as well as studying the link between the two tasks. The best results for (A) showcase the importance of taking into account the sequence-aware modelling of an individual's online shared content, whereas the link between the two tasks has been highlighted on the basis of the best results achieved for (B).

Following last year's setting (Macavaney et al., 2021), we utilised NORC's Enclave. Faced with challenges out of our and NORC's control, we provide directions for shared tasks on sensitive domains (§9). Our aim for the future is to emphasize the need for research on longitudinal tracking and modelling of a user's mental health, under a common experimental setting in a secure environment.

9 Recommendations for the Future

Organising a NLP shared task on highly sensitive datasets is an incredibly challenging effort that relies on the coordination and collaboration of many different actors. In addition to the very useful feedback given by last year's organisers (Macavaney et al., 2021), we have compiled an anonymous feedback questionnaire shared with the 39 members that had access to the DE or were the contact members of a team. In this section, we summarise the key insights gained from the teams' feedback (§9.1) and provide suggestions for future versions of Shared Tasks in such sensitive domains (§9.2).

9.1 Feedback from Participants

The questionnaire consists of 4 multiple choice questions and 2 free-text answers on (Q5) what they liked about this year's shared task vs (Q6) what needs improvement in future editions.

Overview & Q1 – *'My team managed to produce results'*: 18 members completed the feedback form (34% of all 53 participants; 46% of the 39 participants that the questionnaire was shared with), 17 of whom were members of teams that managed to submit their results (within or outside the DE).

Q2 – *'The task description was clear'* (completely disagree to completely agree, [1-5]): All 18 responses were between [3-5], with an average of 4.4/5.0. Based on Q6 shown below, there were two respondents for whom the annotation guidelines and/or resulting labels for Task A were unclear. Providing more examples in such longitudinal tasks from the beginning of the Shared Task can offer an improvement in this regard.

Q3 – *'Communication via slack was easy and efficient.'* (completely disagree to completely agree, [1-5]): Responses were between [3-5], with an average of 4.7/5.0, suggesting that an active communication channel can help participants along the way and is recommended for future editions.

Q4 – *'How was your experience with working on the Data Enclave?'* (5 pre-defined choices): 50% of the respondents said that they faced many difficulties, but would have managed to produce results within the DE nevertheless if there wasn't the major incident during test time (see §4); 4/18 respondents said that there were only some difficulties resulting in minor/medium loss in their productivity. We provide concrete suggestions to this effect in §9.2.

Q5 – *What did you like about the shared task?*: The 17 responses on Q5 can be categorised into two main topics: 13 commented positively on the task itself and 7 on the organisational aspect (quick responses from the organisers – see also Q3 – and working in a secure manner through NORC's DE).

Q6 – *'What did you mostly not like about this year's Shared Task? What issues did you face? How can we improve for the next year?*: Most of the 17 responses concerned issues around working within the DE – from inability to copy/paste to downloading resources. We compile a list of suggestions in §9.2. 2/17 respondents commented on the delay of providing the code (e.g., evaluation, baselines/results); 2/17 commented on the clarity of the annotations (see also Q2); 2/17 also commented on the tightness of deadlines, which were packed towards the end of the Shared Task to allow more time for model training – a wider time frame for future Shared Tasks is recommended. Isolated concerning points (1/17) included the small size of the dataset to reach conclusive outcomes (often a concern in this domain) and inability to perform a direct comparison between systems trained within vs outside the DE (tackled by highlighting the best-performing system for submissions within the DE).

9.2 Suggestions for future organisers

Secure Environment. Given the sensitive nature of data for the Shared Task, it is essential to be able to rely on a secure environment. Following CLPsych 2021, we opted for NORC and their DE. It is important that future organisers plan this collaboration in advance to make sure NORC has sufficient time to identify and secure enough resources and specific expertise to the project. The technical issue faced this year also highlights the need for a wider test-time period, to allow enough time for resolving such cases. Ideally there should be an ongoing collaboration with the DE so that any issues and the necessary expertise to overcome them are built during a sufficiently long period of time.

Libraries and Resources. It is crucial to have a clear pre-defined list of libraries, resources and dependencies (e.g., pre-trained models) that would need to be reviewed before being available in the DE. This means reaching out in advance to the teams and also planning for a trial period of 2 weeks where the teams can access part of the data and check their needs, live. The teams for instance encountered many issues with NLP libraries that required additional downloads of resources when used.[5] It is also important to keep track of the approved/installed libraries each year.

Communication and Peer Support. Following last year's suggestions, we wanted to avoid sending many similar requests to NORC, and try to provide a common setting for people to help each other. We relied on Slack by setting up two dedicated channels, which received very positive feedback and also facilitated the communication between the organisers and NORC. Participants helped each other e.g. in setting up the ssh tunneling for Jupyter Notebook or in identifying the specific issue to report back to NORC (which we have tried to do through a more coordinated effort, where one of the organisers would be the point of contact).

Preparation. Notes from last year's edition already highlighted the complexity of organizing the shared task and recommended more advance planning. Even with that in mind, core challenges remain due to the antithesis between two very different agendas: the intensive experimental work in a very limited time frame (the shared task) and a centralised, step-by-step highly controlled process (the DE). We believe that only through long-term collaboration with DEs such as NORC is it feasible to define a middle-ground working solution which can guarantee high level of security while supporting researchers to develop their solutions. Such collaboration requires the recognition of the importance of DEs by funding bodies and the need to fund long-term collaborations between DEs and research organisations.

[5] e.g., the NLTK tokenizer requires 13MB of Punkt Tokenizer Models, which are not accessible in the DE.

Ethical statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20). Annotators were given contracts and paid fairly in line with University payscales. They were alerted about potentially encountering disturbing content and were advised to take breaks. The annotations are used to train and evaluate natural language processing models for recognising moments of change and linking them to suicidality risk, where the latter is provided by clinical psychology experts. Working with data on online platforms where individuals disclose personal information involves ethical considerations (Mao et al., 2011; Keküllüoğlu et al., 2020). Such considerations include careful analysis and data sharing policies to protect sensitive personal information. Potential risks from the application of NLP models in being able to identify moments of change in individuals' timelines are akin to those in earlier work on personal event identification from social media and the detection of suicidal ideation. Potential mitigation strategies include restricting access to the code base and annotation labels used for evaluation. In this shared task we have asked participants to sign DUA agreements and we opted for a secure data enclave environment to work in.

Acknowledgments

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant ref EP/V030302/1), the Alan Turing Institute (grant ref EP/N510129/1) and especially UKRI funding to promote collaboration between UK and US researchers. Aspects of this work were also supported by an Amazon Research Award and by the National Science Foundation under grant 2124270, and the effort also received internal financial support at NORC. The shared task organizers would like to express their gratitude to the anonymous users of Reddit whose data feature in this year's shared task dataset; to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for TaskB, the American Association of Suicidology; to all participants for their efforts and patience; to the NORC partners and personnel (especially co-author Jeff Leintz, Dariush Wilkowski, Julia Crothers, Bill Olesiuk and the Data Enclave Manager team) for their tremendous contributions and their willingness to put in a great amount of resources in setting up and managing the Enclave and enabling this year's shared task, especially given the short time frame, and finally to NAACL for its support for CLPsych.

References

Falwah AlHamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916.

Diego Arenas, Jon Atkins, Claire Austin, David Beavan, Alvaro Cabrejas Egea, Steven Carlysle-Davies, Ian Carter, Rob Clarke, James Cunningham, Tom Doel, et al. 2019. Design choices for productive, secure, data-intensive research at scale in the cloud. *arXiv preprint arXiv:1908.08737*.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E. Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Michael Barkham, Wolfgang Lutz, and Louis G Castonguay. 2021. *Bergin and Garfield's handbook of psychotherapy and behavior change*. John Wiley & Sons.

Ulya Bayram and Lamia Benhiba. 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Sravani Boinepelli, Shivansh Subramanian, Abhijeeth Singam, Tathagata Raha, and Vasudeva Varma. 2022. Towards capturing changes in mood and identifying suicidality risk. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Ana-Maria Bucur, Hyewon Jang, and Farhana Ferdousi Liza. 2022. Capturing changes in mood over time in longitudinal data using ensemble methodologies. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Prasadith Buddhitha, Ahmed Husseini Orabi, Mahmoud Husseini Orabi, and Diana Inkpen. 2022. Multi-task learning to capture changes in mood over time. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

John Culnan, Damian Y. Romero Diaz, and Steven Bethard. 2022. Exploring transformers and time lag features for predicting changes in mood over time. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Gildo Fabregat, Ander Cejudo, Juan Martinez-Romo, Alicia Pérez, Lourdes Araujo, Nuria Lebeña, Maite Oronoz, and Arantza Casillas. 2022. Approximate nearest neighbour extraction techniques and neural networks for suicide risk prediction in the CLPsych 2022 shared task. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahamanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes C. Eichstaedt, and H. Andrew Schwartz. 2022. WWBP-SQT-lite: Difference embeddings and multi-level models for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Dilara Kekü̈llüoğlu, Walid Magdy, and Kami Vaniea. 2020. Analysing privacy leakage of life events on twitter. In *Proceedings of the 12th ACM Conference on Web Science*.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

Wolfgang Lutz, Torsten Ehrlich, Julian A. Rubel, Nora Hallwachs, Marie-Anna Röttger, Christine Jorasz, Sarah Mocanu, Silja Vocks, Dietmar Schulte, and Armita Tschitsaz-Stucki. 2013. The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, 23:14 – 24.

Sean Macavaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the clpsych 2021 shared task. In *Proc. of CLPsych*, pages 70–80.

Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: An analysis of privacy leaks on twitter. WPES '11, page 1–12, New York, NY, USA. Association for Computing Machinery.

Jonathan G. Shalom and Idan M. Aderka. 2020. A meta-analysis of sudden gains in psychotherapy: Outcome and moderators. *Clinical Psychology Review*, 76:101827.

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.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proc. of ACL*.

Gerrit JJ van den Burg and Christopher KI Williams. 2020. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

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.

A Reddit New: Data Collection

We used the Pushshift API (https://reddit-api.readthedocs.io/en/latest/) to crawl the posts from the following subreddits for Reddit-New: Agoraphobia, HealthAnxiety, autism, hardshipmates, rant, Anxiety, Needafriend, bipolar, lonely, rapecounseling, Anxietyhelp, StopSelfHarm, bipolarreddit, mentalhealth, schizophrenia, BPD, SuicideWatch, bulimia, mentalillness, socialanxiety, COVID19_support, addiction, depression, offmychest, survivorsofabuse, EDAnonymous, adhd, depression_help, panicparty, traumatoolbox, EatingDisorderHope, alcoholism, eating_disorders, psychoticreddit, trueoffmychest, EatingDisorders, anxietysupporters, foreveralone, ptsd, unsentletters.

B Timeline Selection Criteria

When selecting informative timelines, the internal annotators independently classified them into the following categories.

- **Good:** Timelines comprise posts that clearly indicate user mood or at least 1 moment of change in mood.
- **Medium:** Timelines comprise posts from which user mood is challenging to infer. The individual may disclose information about their own life events, but such discussions are objective in tone.
- **Bad:** Timelines comprise posts that do not provide indicators of the user's own mood. If there are posts by the user on subreddits related to mental health, these posts do not clearly relate to the user's own mood (e.g., words of encouragement for other users, cross-posted content shared with intent to help other users rather than themselves).

C Team Selection Assessment Criteria

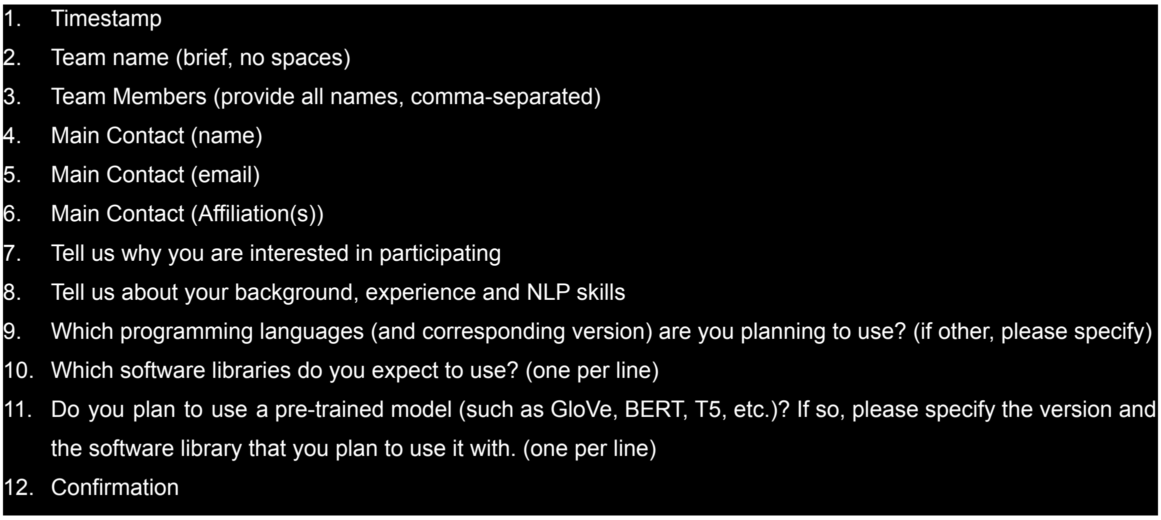
In this section, we outline the assessment criteria used for selecting the teams for participate in the Shared Task. The guidelines were given to two annotators internally, who achieved a high agreement (Pearson correlation ρ=.83).

CLPsych 2022 Shared Tak: Registration of Interest

Guidelines for Reviewing

Aim: We have received applications (Registration of Interest) from 37 teams to participate in the CLPsych Shared Task 2022 (https://clpsych.org/sharedtask2022/). The goal of this reviewing process is to review the submitted applications on the basis of the main questions outlined below.

Registration of Interest Data: Each of the 37 teams that registered their interest provided us with the following information:



1. Timestamp
2. Team name (brief, no spaces)
3. Team Members (provide all names, comma-separated)
4. Main Contact (name)
5. Main Contact (email)
6. Main Contact (Affiliation(s))
7. Tell us why you are interested in participating
8. Tell us about your background, experience and NLP skills
9. Which programming languages (and corresponding version) are you planning to use? (if other, please specify)
10. Which software libraries do you expect to use? (one per line)
11. Do you plan to use a pre-trained model (such as GloVe, BERT, T5, etc.)? If so, please specify the version and the software library that you plan to use it with. (one per line)
12. Confirmation

We anonymised the list presented above and provided you with the following:

1. Number of participants in the team
2. Tell us why you are interested in participating (question 7 form the list above)
3. Tell us about your background, experience and NLP skills (question 8)
4. Which programming languages [...]? (question 9)
5. Which software libraries [...] (question 10)
6. Do you plan to use a pre-trained model [...] (question 11)

The reviewing task will be done solely on the basis of the responses given by each time on questions 2-6. For each team, *please read carefully the responses given by the team to all of the 5 questions prior to assessing their application*. The reason is that even though a reviewing criterion (see below) might seem explicitly related to a particular question (e.g., Criterion 1 seems to be clearly linked to the third question), the responses to the other questions might provide additional information for the team (e.g., the response to the second question might provide you with additional information for Criterion 1).

Assessment Criteria

For each of the three reviewing criteria presented below, please provide your score (half scores, such as “2.5”, are also allowed), your confidence and a justification of your rating.

Criterion 1: Team Background

- Does the background/current work of the team match the requirements of the task? Please rate between 1-5 (half scores allowed):
    - 5: The team has worked/works on similar longitudinal/sequential NLP tasks on mental health.
    - 4: The team has worked/works on similar NLP tasks with a longitudinal or sequential component.
    - 3: The team has worked/works with NLP methods on the mental health domain, though without a sequential/longitudinal component.
    - 2: The team has worked/works with NLP methods, though outside of the mental health domain and without a sequential/longitudinal component .
    - 1: The team has some/no experience with NLP tasks and methods.
- Please justify/comment on your score:
- How confident are you on your assessment?
    - Very
    - Moderately
    - Low

Criterion 2: Commitment

- Based on your assessment, how likely is the team to commit to this task? Please rate between 1-3 (half scores allowed):
    - ○ 3: The task will help the team even to advance their own work, so they are likely to invest a lot of time in the task.
    - ○ 2: The team has shown strong motivation, but their work is not directly linked to the shared task.
    - ○ 1: The team's motivation is not clear/not well explained.
- Please justify/comment on your score:
- How confident are you on your assessment?
    - ○ Very
    - ○ Moderately
    - ○ Low

Criterion 3: Details on Software Requirements

- How detailed are the requests made by the team in terms of software requirements (programming languages & versions, libraries & versions, language models)? Please rate between 1-3 (half scores allowed):
    - ○ 3: The provided information are very detailed. One could set up everything the team has asked for, allowing the team to start working straight away.
    - ○ 2: The provided information are adequate, but not complete. One could probably set up a working environment with many of the required languages/libraries/models, but clarifications would be needed on several aspects (e.g., on specific versions of libraries).
    - ○ 1: The replies of the team are generic/missing. Clarifications are needed in almost all of the requirements.
- Please justify/comment on your score:
- How confident are you on your assessment?
    - ○ Very
    - ○ Moderately
    - ○ Low

Final Question (not part of the assessment): For the isolated participants (i.e., those who are a team on their own: numMembers=1), who should we try to group together so that they form a single team? Try to reply based on their responses to the 5 questions.