Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview)

David E. Losada[1], Fabio Crestani[2], and Javier Parapar[3]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
david.losada@usc.es

[2] Faculty of Informatics,
Universitá della Svizzera italiana (USI), Switzerland
fabio.crestani@usi.ch

[3] Information Retrieval Lab,
Centro de Investigación en Tecnologías de la Información y las Comunicaciones,
Universidade da Coruña, Spain
javierparapar@udc.es

Abstract. This paper provides an overview of eRisk 2020, the fourth edition of this lab under the CLEF conference. The main purpose of eRisk is to explore issues of evaluation methodology, effectiveness metrics and other processes related to early risk detection. Early detection technologies can be employed in different areas, particularly those related to health and safety. This edition of eRisk had two tasks. The first task focused on early detecting signs of self-harm. The second task challenged the participants to automatically filling a depression questionnaire based on user interactions in social media.

1 Introduction

The main purpose of eRisk is to explore issues of evaluation methodologies, performance metrics and other aspects related to building test collections and defining challenges for early risk detection. Early detection technologies are potentially useful in different areas, particularly those related to safety and health. For example, early alerts could be sent when a person starts showing signs of a mental disorder, when a sexual predator starts interacting with a child, or when a potential offender starts publishing antisocial threats on the Internet.

Although the evaluation methodology (strategies to build new test collections, novel evaluation metrics, etc) can be applied on multiple domains, eRisk has so far focused on psychological problems (essentially, depression, self-harm and eating disorders). In 2017 [3, 4], we ran an exploratory task on early detection of depression. This pilot task was based on the evaluation methodology and test collection presented in [2]. In 2018 [6, 5], we ran a continuation of the task on early detection of signs of depression together with a new task on early detection of signs of anorexia. In 2019 [7, 8], we had a a continuation of the task on early detection of signs of anorexia, a new task on early detection of signs of self-harm and a third task oriented to estimate a user's answers to a depression questionnaire based on his interactions on social media.

Over these years, we have been able to compare a number of solutions that employ multiple technologies and models (e.g. Natural Language Processing, Machine Learning, or Information Retrieval). We learned that the interaction between psychological problems and language use is challenging and, in general, the effectiveness of most contributing systems is modest. For example, most challenges had levels of performance (e.g. in terms of F1) below 70%. This suggests that this kind of early prediction tasks require further research and the solutions proposed so far still have much room from improvement.

In 2020, the lab had two campaign-style tasks. The first task had the same orientation of previous early detection tasks. It focused on early detection of signs of self-harm. The second task was a continuation of 2019's third task. It was oriented to analyzing a user's history of posts and extracting useful evidence for estimating the user's depression level. More specifically, the participants had to process the user's posts and, next, estimate the user's answers to a standard depression questionnaire. These tasks are described in the next sections of this overview paper.

2 Task 1: Early Detection of Signs of Self-Harm

This is the continuation of eRisk 2019's T2 task. The challenge consists of sequentially processing pieces of evidence and detect early traces of self-harm as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media. Texts had to be processed in the order they were posted. In this way, systems that effectively perform this task could be applied to sequentially monitor user interactions in blogs, social networks, or other types of online media.

The test collection for this task had the same format as the collection described in [2]. The source of data is also the same used for previous eRisks. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, self-harm and non-self-harm, and, for each user, the collection contains a sequence of writings (in chronological order).

In 2019, we moved from a chunk-based release of data (used in 2017 and 2018) to a item-by-item release of data. We set up a server that iteratively gave user writings to the participating teams. In 2020, the same server was used to provide the users' writings during the test stage. More information about the server can be found at the lab website[4].

[4] http://early.irlab.org/server.html

Table 1. Task1 (self-harm). Main statistics of the train and test collections

	Train		Test	
	Self-Harm	*Control*	*Self-Harm*	*Control*
Num. subjects	41	299	104	319
Num. submissions (posts & comments)	6,927	163,506	11,691	91,136
Avg num. of submissions per subject	169.0	546.8	112.4	285.6
Avg num. of days from first to last submission	≈ 495	≈ 500	≈ 270	≈ 426
Avg num. words per submission	24.8	18.8	21.4	11.9

The 2020 task was organized into two different stages:

- Training stage. Initially, the teams that participated in this task had access to a training stage where we released the whole history of writings for a set of training users (we provided all writings of all training users), and we indicated what users had explicitly mentioned that they have done self-harm. The participants could therefore tune their systems with the training data. In 2020, the training data for Task 1 was composed of all 2019's T2 users.
- Test stage. The test stage consisted of a period of time where the participants had to connect to our server and iteratively got user writings and sent responses. Each participant had the opportunity to stop and make an alert at any point of the user chronology. After reading each user post, the teams had to choose between: i) emitting an alert on the user, or ii) making no alert on the user. Alerts were considered as final (i.e. further decisions about this individual were ignored), while *no alerts* were considered as non-final (i.e. the participants could later submit an alert for this user if they detected the appearance of risk signs). This choice had to be made for each user in the test split. The systems were evaluated based on the accuracy of the decisions and the number of user writings required to take the decisions (see below). A REST server was built to support the test stage. The server iteratively gave user writings to the participants and waited for their responses (no new user data provided until the system said alert/no alert). This server was running from March 2nd, 2020 to May 24th, 2020[5].

Table 1 reports the main statistics of the train and test collections used for T1. Evaluation measures are discussed in the next section.

2.1 Decision-based Evaluation

This form of evaluation revolves around the (binary) decisions taken for each user by the participating systems. Besides standard classification measures (Precision, Recall and F1[6]), we computed $ERDE$, the early risk detection error used in the previous editions of the lab. A full description of $ERDE$ can be found in [2].

[5] In the initial configuration, the test period was shorter but, because of the COVID-19 situation, we decided to extend the test stage in order to facilitate participation.

[6] computed with respect to the positive class.

Essentially, $ERDE$ is an error measure that introduces a penalty for late correct alerts (true positives). The penalty grows with the delay in emitting the alert, and the delay is measured here as the number of user posts that had to be processed before making the alert.

Since 2019, we complemented the evaluation report with additional decision-based metrics that try to capture additional aspects of the problem. These metrics try to overcome some limitations of $ERDE$, namely:

- the penalty associated to true positives goes quickly to 1. This is due to the functional form of the cost function (sigmoid).
- a perfect system, which detects the true positive case right after the first round of messages (first chunk), does not get error equal to 0.
- with a method based on releasing data in a chunk-based way (as it was done in 2017 and 2018) the contribution of each user to the performance evaluation has a large variance (different for users with few writings per chunk vs users with many writings per chunk).
- $ERDE$ is not interpretable.

Some research teams have analysed these issues and proposed alternative ways for evaluation. Trotzek and colleagues [10] proposed $ERDE_o^{\%}$. This is a variant of ERDE that does not depend on the number of user writings seen before the alert but, instead, it depends on the *percentage* of user writings seen before the alert. In this way, user's contributions to the evaluation are normalized (currently, all users weight the same). However, there is an important limitation of $ERDE_o^{\%}$. In real life applications, the overall number of user writings is not known in advance. Social Media users post contents online and screening tools have to make predictions with the evidence seen. In practice, you do not know when (and if) a user's thread of message is exhausted. Thus, the performance metric should not depend on such lack of knowledge about the total number of user writings.

Another proposal of an alternative evaluation metric for early risk prediction was done by Sadeque and colleagues [9]. They proposed $F_{latency}$, which fits better with our purposes. This measure is described next.

Imagine a user $u \in U$ and an early risk detection system that iteratively analyzes u's writings (e.g. in chronological order, as they appear in Social Media) and, after analyzing k_u user writings ($k_u \geq 1$), takes a binary decision $d_u \in \{0, 1\}$, which represents the decision of the system about the user being a risk case. By $g_u \in \{0, 1\}$, we refer to the user's golden truth label. A key component of an early risk evaluation should be the delay on detecting true positives (we do not want systems to detect these cases too late). Therefore, a first and intuitive measure of delay can be defined as follows[7]:

[7] Observe that Sadeque et al (see [9], pg 497) computed the latency for all users such that $g_u = 1$. We argue that latency should be computed only for the true positives. The false negatives ($g_u = 1$, $d_u = 0$) are not detected by the system and, therefore, they would not generate an alert.

$$\text{latency}_{TP} = \text{median}\{k_u : u \in U, d_u = g_u = 1\} \tag{1}$$

This measure of latency goes over the true positives detected by the system and assesses the system's delay based on the median number of writings that the system had to process to detect such positive cases. This measure can be included in the experimental report together with standard measures such as Precision (P), Recall (R) and the F-measure (F):

$$P = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : d_u = 1|} \tag{2}$$

$$R = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : g_u = 1|} \tag{3}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

Furthermore, Sadeque et al. proposed a measure, $F_{latency}$, which combines the effectiveness of the decision (estimated with the F measure) and the delay[8]. This is based on multiplying F by a penalty factor based on the median delay. More specifically, each individual (true positive) decision, taken after reading k_u writings, is assigned the following penalty:

$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}} \tag{5}$$

where p is a parameter that determines how quickly the penalty should increase. In [9], p was set such that the penalty equals 0.5 at the median number of posts of a user[9]. Observe that a decision right after the first writing has no penalty ($penalty(1) = 0$). Figure 1 plots how the latency penalty increases with the number of observed writings.

The system's overall speed factor is computed as:

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\}) \tag{6}$$

speed equals 1 for a system whose true positives are detected right at the first writing. A slow system, which detects true positives after hundreds of writings, will be assigned a speed score near 0.

Finally, the *latency-weighted* F score is simply:

$$F_{latency} = F \cdot speed \tag{7}$$

[8] Again, we adopt Sadeque et al.'s proposal but we estimate latency only over the true positives.

[9] In the evaluation we set p to 0.0078, a setting obtained from the eRisk 2017 collection.

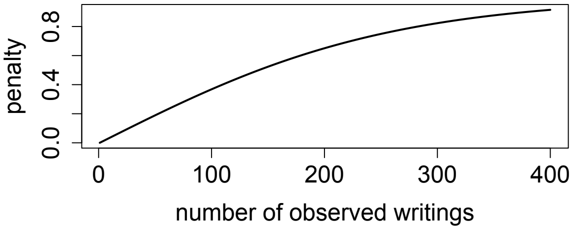


Fig. 1. Latency penalty increases with the number of observed writings (k_u)

Since 2019 user's data was processed by the participants in a post by post basis (i.e. we avoided a chunk-based release of data). Under these conditions, the evaluation approach has the following properties:

- smooth grow of penalties.
- a perfect system gets $F_{latency} = 1$.
- for each user u the system can opt to stop at any point k_u and, therefore, now we do not have the effect of an imbalanced importance of users.
- $F_{latency}$ is more interpretable than $ERDE$.

2.2 Ranking-based Evaluation

This section discusses an alternative form of evaluation, which was used as a complement of the evaluation described above. After each release of data (new user writing) the participants had to send back the following information (for each user in the collection): i) a decision for the user (alert/no alert), which was used to compute the decision-based metrics discussed above, and ii) a score that represents the user's level of risk (estimated from the evidence seen so far). We used these scores to build a ranking of users in decreasing estimation of risk. For each participating system, we have one ranking at each point (i.e., ranking after 1 writing, ranking after 2 writings, etc.). This simulates a continuous re-ranking approach based on the evidence seen so far. In a real life application, this ranking would be presented to an expert user who could take decisions (e.g. by inspecting the rankings).

Each ranking can be scored with standard IR metrics, such as P@10 or NDCG. We therefore report the ranking-based performance of the systems after seeing k writings (with varying k).

2.3 Task 1: Results

Table 6 shows the participating teams, the number of runs submitted and the approximate lapse of time from the first response to the last response. This lapse of time is indicative of the degree of automation of each team's algorithms. A few of the submitted runs processed the entire thread of messages (nearly 2000), but many variants opted for stopping earlier. Six teams processed the thread of messages in a reasonably fast way (less than a day or so for processing the entire history of user messages). The rest of the teams took several days to run the whole process. Some teams took even more than a week. This suggests that they incorporated some form of offline processing.

Table 3 reports the decision-based performance achieved by the participating teams.

In terms of Precision, $F1$, ERDE measures and latency-weighted $F1$, the best performing runs were submitted by the iLab team. The first two iLab runs had extremely high precision (.833 and .913, respectively) and the first one (run #0) had the highest latency-weighted F1 (.658). These runs had low levels of recall (.577 and .404) and they only analyzed a median of 10 user writings. This suggests that you can get to a reasonably high level of precision based on a few user writings. The main limitation of these best performing runs is the low levels of recall achieved. In terms of $ERDE$, the best performing runs show low levels of error (.134 and .071). ERDE measures set a strong penalty on late decisions and the two best runs show a good balance between the accuracy of the decisions and the delays (latency of the true positives was 2 and 45, respectively, for the two runs that achieved the lowest $ERDE_5$ and $ERDE_{50}$).

Other teams submitted high recall runs but their precision was very low and, thus, these automatic methods are hardly usable to filter out non-risk cases.

Most teams submitted quick decisions. Only iLab and prhlt-upv have some runs that analysed more than a hundred submissions before emitting the alerts (mean latencies higher than 100).

Overall, these results suggest that with a few dozen user writings some systems led to reasonably high effectiveness. The best predictive algorithms could be used to support expert humans in early detecting signs of self-harm.

Table 4 reports the ranking-based performance achieved by the participating teams. Some teams only processed a few dozens of user writings and, thus, we could only compute their rankings of users for the initial points.

Some teams (e.g., INAOE-CIMAT or BioInfo@UAVR) have the same levels of ranking-based effectiveness over multiple points (after 1 writing, after 100 writings, and so forth). This suggests that these teams did not change the risk scores estimated from the initial stages (or their algorithms were not able to enhance their estimations as more evidence was seen).

Other participants (e.g., EFE, iLab or hildesheim) behave as expected: the rankings of estimated risk get better as they are built from more user evidence. Notably, some iLab variants variants led to almost perfect P@10 and $NDCG$@10 performance after analyzing more than 100 writings. The $NDCG$@100 scores achieved by this team after 100 or 500 writings were also quite effective (above .81 for all variants). This suggests that, with enough pieces of evidence, the methods implemented by this team are highly effective at prioritizing at-risk users.

Table 2. Task 1. Participating teams: number of runs, number of user writings processed by the team, and lapse of time taken for the whole process.

team	#runs	#user writings	lapse of time
		processed	(from 1st to last response)
UNSL	5	1990	10 hs
INAOE-CIMAT	5	1989	7 days + 7 hs
BiTeM	5	1	1 min
EFE	3	1991	12 hs
NLP-UNED	5	554	1 day
BioInfo@UAVR	3	565	2 days + 21 hs
SSN_NLP	5	222	3 hs
Anji	5	1990	1 day + 3 hs
Hildesheim	5	522	72 days + 20 hs
RELAI	5	1990	2 days + 8 hs
Prhlt-upv	5	627	1 day + 8 hs
iLab	5	954	20 hs

3 Task 2: Measuring the Severity of the Signs of Depression

This task is a continuation of 2019's T3 task. The task consists of estimating the level of depression from a thread of user submissions. For each user, the participants were given the user's full history of postings (in a single release of data) and the participants had to fill a standard depression questionnaire based on the evidence found in the history of postings. In 2020, the participants had the opportunity to use 2019's data as training data (filled questionnaires and SM submissions from the 2019 users, i.e. a training set composed of 20 users).

The questionnaires are derived from the Beck's Depression Inventory (BDI)[1], which assesses the presence of feelings like sadness, pessimism, loss of energy, etc, for the detection of depression. The questionnaire contains 21 questions (see figs 2, 3).

The task aims at exploring the viability of automatically estimating the severity of the multiple symptoms associated with depression. Given the user's history of writings, the algorithms had to estimate the user's response to each individual question. We collected questionnaires filled by Social Media users together with their history of writings (we extracted each history of writings right after the user provided us with the filled questionnaire). The questionnaires filled by the users (ground truth) were used to assess the quality of the responses provided by the participating systems.

The participants were given a dataset with 70 users and they were asked to produce a file with the following structure:

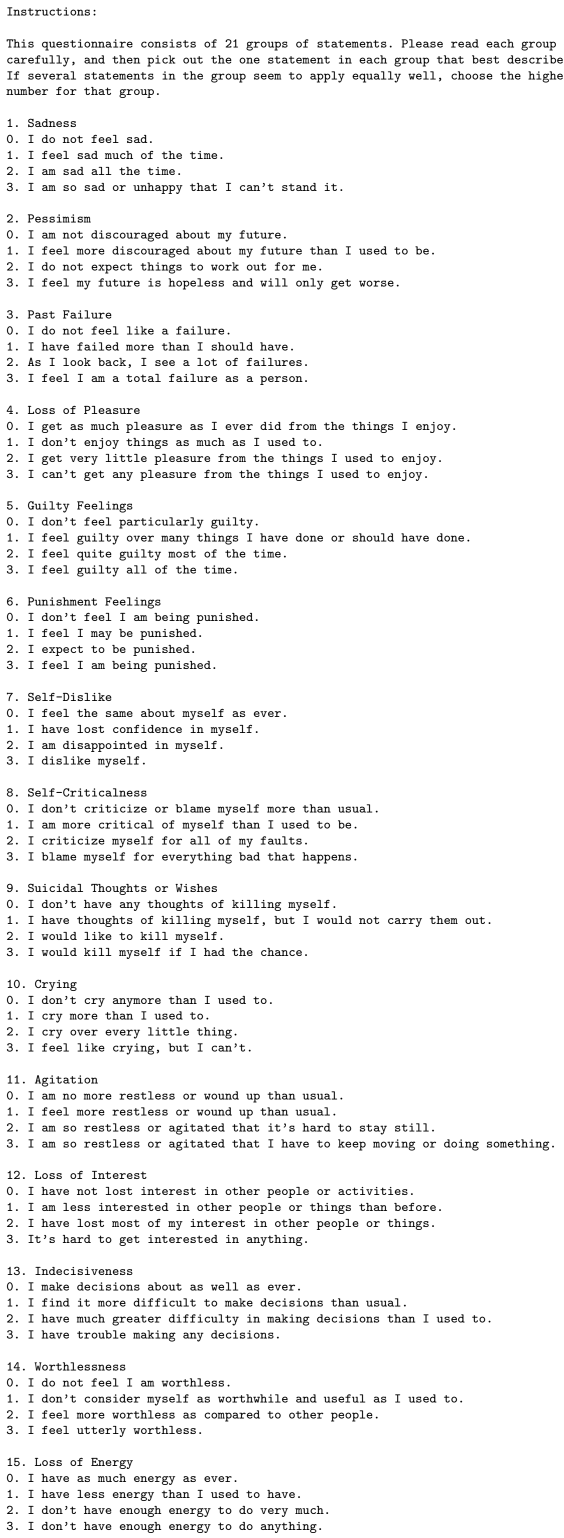
```
username1 answer1 answer2 .... answer21
username2 ....
....
```

Table 3. Task 1. Decision-based evaluation

team name	run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted$
	id								$F1$
Hildesheim	0	.248	**1**	.397	.292	.196	1	1	.397
Hildesheim	1	.246	**1**	.395	.304	.185	5	.984	.389
Hildesheim	2	.297	.740	.424	.237	.226	1	1	.424
Hildesheim	3	.270	.942	.420	.400	.251	33.5	.874	.367
Hildesheim	4	.256	.990	.406	.409	.210	12	.957	.389
UNSL	0	.657	.423	.515	.191	.155	2	.996	.513
UNSL	1	.618	.606	.612	.172	.124	2	.996	.609
UNSL	2	.606	.548	.576	.267	.142	11	.961	.553
UNSL	3	.598	.529	.561	.267	.149	12	.957	.537
UNSL	4	.545	.519	.532	.271	.151	12	.957	.509
EFE	0	.730	.519	.607	.257	.142	11	.961	.583
EFE	1	.625	.625	.625	.268	.117	11	.961	.601
EFE	2	.496	.615	.549	.283	.140	11	.961	.528
iLab	0	.833	.577	.682	.252	.111	10	.965	**.658**
iLab	1	**.913**	.404	.560	.248	.149	10	.965	.540
iLab	2	.544	.654	.594	**.134**	.118	2	.996	.592
iLab	3	.564	.885	.689	.287	**.071**	45	.830	.572
iLab	4	.828	.692	**.754**	.255	.255	100	.632	.476
prhlt-upv	0	.469	.654	.546	.291	.154	41	.845	.462
prhlt-upv	1	.710	.212	.326	.251	.235	133	.526	.172
prhlt-upv	2	.271	.577	.369	.339	.269	51.5	.806	.298
prhlt-upv	3	.846	.212	.338	.248	.232	133	.526	.178
prhlt-upv	4	.765	.375	.503	.253	.194	42	.841	.423
INAOE-CIMAT	0	.488	.567	.524	.203	.145	4	.988	.518
INAOE-CIMAT	1	.500	.548	.523	.193	.144	4	.988	.517
INAOE-CIMAT	2	.848	.375	.520	.207	.160	5	.984	.512
INAOE-CIMAT	3	.525	.702	.601	.174	.119	3	.992	.596
INAOE-CIMAT	4	.788	.394	.526	.198	.160	4	.988	.519
BioInfo@UAVR	0	.609	.375	.464	.260	.178	14	.949	.441
BioInfo@UAVR	1	.591	.654	.621	.273	.120	11	.961	.597
BioInfo@UAVR	2	.629	.375	.470	.259	.177	13	.953	.448
RELAI	0	.341	.865	.489	.188	.136	2	.996	.487
RELAI	1	.350	.885	.501	.190	.130	2	.996	.499
RELAI	2	.438	.740	.550	.245	.132	8	.973	.535
RELAI	3	.291	.894	.439	.306	.168	7	.977	.428
RELAI	4	.381	.846	.525	.260	.141	7	.977	.513
SSN_NLP	0	.264	**1**	.419	.206	.170	1	1.0	.419
SSN_NLP	1	.283	**1**	.442	.205	.158	1	1.0	.442
SSN_NLP	2	.287	.990	.445	.228	.159	2	.996	.443
SSN_NLP	3	.688	.423	.524	.233	.171	15.5	.944	.494
SSN_NLP	4	.287	.952	.441	.263	.214	4	.988	.436
BiTeM	0	.333	.01	.02	.245	.245	1	1.0	.019
BiTeM	1	0	0	0					
BiTeM	2	0	0	0					
BiTeM	3	0	0	0					
BiTeM	4	0	0	0					
NLP-UNED	0	.237	.913	.376	.423	.199	11	.961	.362
NLP-UNED	1	.246	**1**	.395	.210	.185	1	1.0	.395
NLP-UNED	2	.246	**1**	.395	.210	.185	1	1.0	.395
NLP-UNED	3	.246	**1**	.395	.210	.185	1	1.0	.395
NLP-UNED	4	.246	**1**	.395	.210	.185	1	1.0	.395
Anji	0	.266	**1**	.420	.205	.167	1	1.0	.420
Anji	1	.266	**1**	.420	.211	.167	1	1.0	.420
Anji	2	.269	**1**	.424	.213	.164	1	1.0	.424
Anji	3	.333	.038	.069	.248	.243	7	.977	.067
Anji	4	.258	.990	.410	.208	.174	1	1.0	.410

Table 4. Task 1. Ranking-based evaluation

		1 writing			100 writings			500 writings			1000 writings		
team	run	P	$NDCG$	$NDCG$	P	$NDCG$	$NDCG$	P	$NDCG$	$NDCG$	P	$NDCG$	$NDCG$
		@10	@10	@100	@10	@10	@100	@10	@10	@100	@10	@10	@100
Hildesheim	0	.1	.10	.26	.4	.43	.42	.5	.53	.42			
Hildesheim	1	.4	.44	.30	.5	.48	.49	.5	.54	.57			
Hildesheim	2	.2	.15	.24	**1**	**1**	.69	**1**	**1**	.68			
Hildesheim	3	.2	.14	.20	.1	.07	.13	.1	.06	.11			
Hildesheim	4	.2	.16	.18	**1**	**1**	.62	**1**	**1**	.69			
UNSL	0	**.9**	.92	.47	**1**	**1**	.60	**1**	**1**	.60	**1**	**1**	.60
UNSL	1	.8	.87	.55	**1**	**1**	.76	**1**	**1**	.75	**1**	**1**	.75
UNSL	2	.7	.80	.42	.8	.84	.70	.8	.87	.74	.9	.94	.73
UNSL	3	.7	.79	.43	.8	.84	.70	.8	.87	.74	.9	.94	.73
UNSL	4	.5	.63	.36	.8	.86	.62	.8	.86	.62	.8	.86	.62
EFE	0	.7	.65	.59	**1**	**1**	.78	**1**	**1**	.79	**1**	**1**	.79
EFE	1	.6	.54	.58	**1**	**1**	.78	**1**	**1**	.80	**1**	**1**	**.80**
EFE	2	.6	.64	.55	.9	.92	.71	.9	.92	.73	.9	.92	.72
iLab	0	.8	.88	.63	**1**	**1**	.82	**1**	**1**	.83			
iLab	1	.7	.69	.60	**1**	**1**	.82	.9	.94	.81			
iLab	2	.7	.69	.60	**1**	**1**	.82	.9	.94	.81			
iLab	3	**.9**	**.94**	**.66**	**1**	**1**	**.83**	**1**	**1**	**.84**			
iLab	4	.8	.88	.63	**1**	**1**	.82	**1**	**1**	.83			
prhlt-upv	0	.2	.13	.30	.9	.93	.68	**1**	**1**	.68			
prhlt-upv	1	**.9**	.90	.63	.9	.92	.70	.9	.81	.75			
prhlt-upv	2	.5	.41	.42	.6	.69	.48	.6	.69	.48			
prhlt-upv	3	**.9**	.90	.63	.9	.92	.70	.9	.81	.75			
prhlt-upv	4	.8	.75	.49	**1**	**1**	.70	.9	.90	.69			
INAOE-CIMAT	0	.3	.25	.30	.3	.26	.24	.3	.26	.24	.3	.26	.24
INAOE-CIMAT	1	.3	.25	.30	.3	.26	.24	.3	.26	.24	.3	.26	.24
INAOE-CIMAT	2	.3	.25	.30	.3	.26	.24	.3	.26	.24	.3	.26	.24
INAOE-CIMAT	3	.3	.25	.30	.3	.26	.24	.3	.26	.24	.3	.26	.24
INAOE-CIMAT	4	.3	.25	.30	.3	.26	.24	.3	.26	.24	.3	.26	.24
BioInfo@UAVR	0	.6	.62	.33	.6	.62	.31	.6	.62	.31			
BioInfo@UAVR	1	.6	.62	.33	0	0	.07	0	0	.04			
BioInfo@UAVR	2	.6	.62	.33	.6	.62	.31	.6	.62	.31			
RELAI	0	.7	.80	.52	.8	.87	.52	.8	.87	.52	.8	.87	.50
RELAI	1	.3	.28	.43	.6	.69	.47	.6	.69	.47	.7	.75	.47
RELAI	2	.2	.20	.27	.7	.81	.63	.8	.87	.70	.8	.87	.72
RELAI	3	.2	.20	.27	.9	.94	.51	**1**	**1**	.59	**1**	**1**	.60
RELAI	4	.2	.20	.27	.7	.68	.59	**1**	**1**	.71	.9	.81	.66
SSN_NLP	0	.7	.68	.50	.5	.38	.43						
SSN_NLP	1	.7	.68	.50	.5	.38	.43						
SSN_NLP	2	.7	.68	.50	.5	.38	.43						
SSN_NLP	3	0	0	.22	.1	.12	.16						
SSN_NLP	4	.7	.68	.50	.5	.38	.43						
BiTeM	0												
BiTeM	1												
BiTeM	2												
BiTeM	3												
BiTeM	4												
NLP-UNED	0	.7	.69	.49	.6	.73	.26	.6	.73	.24			
NLP-UNED	1	.6	.62	.27	.2	.27	.18	.2	.27	.16			
NLP-UNED	2	.6	.62	.27	.2	.27	.18	.2	.27	.16			
NLP-UNED	3	.6	.62	.27	.2	.27	.18	.2	.27	.16			
NLP-UNED	4	.6	.62	.27	.2	.27	.18	.2	.27	.16			
Anji	0	.7	.73	.58	.6	.57	.46	.4	.32	.36	.4	.32	.36
Anji	1	**.9**	.81	.54	.8	.62	.69	.8	.62	.70	.8	.62	.69
Anji	2	.8	.88	.51	.7	.76	.58	.5	.34	.47	.6	.48	.50
Anji	3	.3	.25	.31	.3	.28	.27	.3	.26	.27	.3	.26	.27
Anji	4	.3	.22	.25	.6	.44	.59	.6	.44	.61	.6	.44	.60



```
Instructions:

This questionnaire consists of 21 groups of statements. Please read each group of statements
carefully, and then pick out the one statement in each group that best describes the way you feel.
If several statements in the group seem to apply equally well, choose the highest
number for that group.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can’t stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don’t enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can’t get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don’t feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

6. Punishment Feelings
0. I don’t feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.

7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.

8. Self-Criticalness
0. I don’t criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes
0. I don’t have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying
0. I don’t cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can’t.

11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it’s hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It’s hard to get interested in anything.

13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness
0. I do not feel I am worthless.
1. I don’t consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don’t have enough energy to do very much.
3. I don’t have enough energy to do anything.
```

Fig. 2. Beck’s Depression Inventory (part 1)

```
16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
la. I sleep somewhat more than usual.
lb. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a Iot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can’t get back to sleep.

17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.

18. Changes in Appetite
0. I have not experienced any change in my appetite.
la. My appetite is somewhat less than usual.
lb. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.

19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can’t concentrate as well as usual.
2. It’s hard to keep my mind on anything for very long.
3. I find I can’t concentrate on anything.

20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely
```

Fig. 3. Beck’s Depression Inventory (part 2)

Each line has a user identifier and 21 values. These values correspond to the responses to the questions of the depression questionnaire (the possible values are 0, 1a, 1b, 2a, 2b, 3a, 3b -for questions 16 and 18- and 0, 1, 2, 3 -for the rest of the questions-).

3.1 Task 2: Evaluation Metrics

For consistency purposes, we employed the same evaluation metrics utilised in 2019. These metrics assess the quality of a questionnaire filled by a system in comparison with the real questionnaire filled by the actual Social Media user:

- **Average Hit Rate** (AHR): Hit Rate (HR) averaged over all users. HR is a stringent measure that computes the ratio of cases where the automatic questionnaire has exactly the same answer as the real questionnaire. For example, an automatic questionnaire with 5 matches gets HR equal to 5/21 (because there are 21 questions in the form).
- **Average Closeness Rate** (ACR): Closeness Rate (CR) averaged over all users. CR takes into account that the answers of the depression questionnaire represent an ordinal scale. For example, consider the #17 question:

```
17. Irritability
0. I am no more irritable than usual.
```

```
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.
```

Imagine that the real user answered "0". A system S1 whose answer is "3" should be penalised more than a system S2 whose answer is "1".
For each question, CR computes the absolute difference (ad) between the real and the automated answer (e.g. ad=3 and ad=1 for S1 and S2, respectively) and, next, this absolute difference is transformed into an effectiveness score as follows: $CR = (mad - ad)/mad$, where mad is the maximum absolute difference, which is equal to the number of possible answers minus one.
NOTE: in the two questions (#16 and #18) that have seven possible answers $\{0, 1a,\ 1b,\ 2a,\ 2b, 3a\ , 3b\}$ the pairs $(1a, 1b)$, $(2a, 2b)$, $(3a, 3b)$ are considered equivalent because they reflect the same depression level. As a consequence, the difference between $3b$ and 0 is equal to 3 (and the difference between $1a$ and $1b$ is equal to 0).

- **Average DODL (ADODL)**: Difference between overall depression levels (DODL) averaged over all users. The previous measures assess the systems' ability to answer each question in the form. DODL, instead, does not look at question-level hits or differences but computes the overall depression level (sum of all the answers) for the real and automated questionnaire and, next, the absolute difference ($ad\_overall$) between the real and the automated score is computed.
  Depression levels are integers between 0 and 63 and, thus, DODL is normalised into [0,1] as follows: $DODL = (63 - ad\_overall)/63$.
- **Depression Category Hit Rate (DCHR)**. In the psychological domain, it is customary to associate depression levels with the following categories:

```
minimal depression (depression levels 0-9)
mild depression (depression levels 10-18)
moderate depression (depression levels 19-29)
severe depression (depression levels 30-63)
```

The last effectiveness measure consists of computing the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire.

3.2 Task 2: Results

Table 5 presents the results achieved by the participants in this task. To put things in perspective, the table also reports (lower block) the performance achieved by three baseline variants: all 0s and all 1s, which consist of sending the same response (0 or 1) for all the questions, and random, which is the average performance (averaged over 1000 repetitions) achieved by an algorithm that randomly chooses among the possible answers.

Although the teams could use training data from 2019 (while 2019's participants had no training data), the performance scores tend to be lower than 2019's

Table 5. Task 2. Performance Results

Run	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	**38.30%**	69.21%	76.01%	30.00%
iLab run1	36.73%	68.68%	81.07%	27.14%
iLab run2	37.07%	**69.41%**	81.70%	27.14%
iLab run3	35.99%	69.14%	82.93%	34.29%
prhlt_logreg_features	34.01%	67.07%	80.05%	**35.71%**
prhlt_svm_use	36.94%	69.02%	81.72%	31.43%
prhlt_svm_features	34.56%	67.44%	80.63%	**35.71%**
svm_features	34.56%	67.44%	80.63%	**35.71%**
relai_context_paral_user	36.80%	68.37%	80.84%	22.86%
relai_context_sim_answer	21.16%	55.40%	73.76%	27.14%
relai_lda_answer	28.50%	60.79%	79.07%	30.00%
relai_lda_user	36.39%	68.32%	**83.15%**	34.29%
relai_sylo_user	37.28%	68.37%	80.70%	20.00%
Run1_resultat_CNN_Methode_max	34.97%	67.19%	76.85%	25.71%
Run2_resultat_CNN_Methode_suite	32.79%	66.08%	76.33%	17.14%
Run3_resultat_BILSTM_Methode_max	34.01%	67.78%	79.30%	22.86%
Run4_resultat_BILSTM_Methode_suit	33.54%	67.26%	78.91%	20.00%
all 0s	36.26%	64.22%	64.22%	14.29%
all 1s	29.18%	73.38%	81.95%	25.71%
random (avg 1000 repetitions)	23.94%	58.44%	75.22%	26.53%

performance scores (only ADODL had higher performance). This could be due to various reasons, including the intrinsic difficulty of the task and the lack of discussion on SM of psychological concerns by 2020 users.

In terms of AHR, the best performing run (BioInfo@UAVR) only got 38.30% of the answers right. The scores of the distance-based measure (ACR) are below 70%. Most of the questions have four possible answers and, thus, a random algorithm would get AHR near 25%[10]. This suggests that the analysis of the user posts was useful at extracting some signals or symptoms related to depression. However, ADODL and, particularly, DCHR show that the participants, although effective at answering some depression-related questions, do not fare well at estimating the overall level of depression of the individuals. For example, the best performing run gets the depression category right for only 35.71% of the individuals.

Overall, these experiments indicate that we are still far from a really effective depression screening tool. In the near future, it will be interesting to further analyze the participants' estimations in order to investigate which particular BDI questions are easier or harder to automatically answer based on Social Media activity.

[10] Actually, slightly less than 25% because a couple of questions have more than four possible answers.

4 Participating Teams

Table 6 reports the participating teams and the runs that they submitted for each eRisk task. The next paragraphs give a brief summary on the techniques implemented by each of them. Further details are available at the CLEF 2020 working notes proceedings.

Table 6. eRisk 2020. Participants

	T1	T2
team	#runs	#runs
UNSL	5	
INAOE-CIMAT	5	
BiTeM	5	
EFE	3	
NLP-UNED	5	
BioInfo@UAVR	3	1
SSN_NLP	5	
Anji	5	
Hildesheim	5	
RELAI	5	5
Prhlt-upv	5	4
iLab	5	3
USDB		4

EFE. This is a team from Dept. of Computer Engineering, Ferdowsi University of Mashhad (Iran). They implemented three variants for Task 1 that represent the texts using Word2Vec representations and performed experiments using Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and Support Vector Machines (SVMs). The entire system is an ensemble multi-level method based on SVM, CNN, and LSTM, which are fine-tuned by attention layers.

USDB. This is a joint collaboration between the LRDSI Laboratory (BLIDA 1 University, Algeria) and the Information Assurance and Security Research Group (Faculty of Computing, Universiti Teknologi Malaysia, Malaysia). This team participated in Task 2 and transformed the user's texts into distributed representations following the Skip-gram model. Next, sentences are encoded using a CNN or Bi-LSTM model (or with Recurrent Neural Networks and Long Bi-LSTM). For each user post, the models generate 21 outputs, which are the answers to the BDI questions. Finally, the user's overall questionnaire is obtained by selecting, for each BDI question, the most frequent answer.

NLP-UNED. This is a joint effort by the NLP & IR Group, at Universidad Nacional de Educación a Distancia (UNED), Spain and the Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS), Spain. These researchers, which participated in Task 1, implemented a machine learning approach using textual features and a SVM classifier. In order to extract relevant features, this team followed a sliding window approach that handles the last messages published by any given user. The features considered a wide range of variables, such as title length, words in the title, punctuation, emoticons, and other feature sets obtained from sentiment analysis, first person pronouns, and NSSI words.

Hildesheim. This team, from the Institute for Information Science and Natural Language Processing (University of Hildesheim, Germany), implemented four variants that apply different methods for Task 1 and a fifth ensemble system that combines the four variants. The four methods utilize different types of features, such as time intervals between posts, and the sentiment and semantics of the writings. To this aim, a neural network approach using bag-of-words vectors and contextualized word embeddings was employed.

BioInfo@UAVR. This team comes from the Bioinformatics group of the Institute of Electronics and Engineering Informatics (University of Aveiro, Portugal). They participated in both tasks. Their approach built upon the algorithms proposed by them for eRisk 2019. For Task 1, they considered a bag of words approach with tf-idf features and employed linear Support Vector Machines with Stochastic Gradient Descent and Passive Aggressive classifiers. For Task 2, the method is based on training a machine learning model using an external dataset and, next, employing the learnt classifier against the eRisk 2020 data. These authors considered psycholinguistic and behavioural features in their attempt to associate the BDI responses with the user's posts.

INAOE-CIMAT. This is a joint effort by Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico and Centro de Investigación en Matemáticas (CIMAT), Mexico. This team participated in the first task and proposed a so-called *Bag of Sub-Emotions* approach that represents the posts of the users using a set of sub-emotions. This representation, which was subsequently combined with bag of words representations, captures the emotions and topics that users with signs of self-harm tend to use. At test time, they experimented with five variants that estimate the temporal stability associated to the users' posts.

SSN-NLP. These participants come from the SSN College Of Engineering, Chennai (India) and participated in Task 1. Given the training data, they experimented with five alternative classification approaches (using tf/idf representations and Bernoulli Naive Bayes, Gradient Boosting, Random Forest, or Extra Trees; or CNNs together with Word2Vec representations).

iLab. This is a joint collaboration between Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain and the Department of Computer and Information Sciences, University of Strathclyde, UK. This team participated in both tasks. They used BERT-based classifiers which were trained specifically for each task. A variety of pretrained models were tested against training data (including BERT, DistillBERT, RoBERTa and XLM-RoBERTa). Rather than using the task's training data, these participants created four new training datasets from Reddit. The submitted runs for Task 1 were based on XLM-RoBERTa. For Task 2, they employed similar methods as the ones employed for Task 1, but they treated the problem as a multi-class labelling problem (one problem for each BDI question).

Prhlt-upv. This team is composed of researchers from Universitat Politécnica de Valencia and from University of Bucharest. They employed multidimensional representations of language and deep learning models (including hierarchical architectures, pre-trained transformers and language models). This team participated in both tasks. For Task 1, they utilized content features, style features, LIWC features, emotions and sentiments. Different strategies were implemented to represent the users' submissions (e.g., augmenting the data by sampling from the user's history or computing a rolling average associated to the most recent estimations). For Task 2, they employed simpler learning models (SVMs and logistic regression) and some of the features extracted for Task 1. The problem was tackled as a multi-label multi-class problem where one model was trained for each BDI question.

Relai. This team comes from University of Quebec in Montreal, Canada. These researchers participated in both tasks, and addressed them using topic modeling al- gorithms (LDA and Anchor Variant), neural models with three different architectures (Deep Averaging Networks, Contextualizers, and RNNs), and an approach based on writing styles. Some of the variants considered stylometry variables, such as Part-of-Speech, frequent n-grams, punctuation, length of words/sentences and usage of uppercase or hyperlinks.

5 Conclusions

This paper provided an overview of eRisk 2020. This was the fourth edition of this lab and the lab's activities concentrated on two different types of tasks: early detection of signs of self-harm (T1), where the participants had a sequential access to the user's social media posts and they had to send alerts about at-risk individuals, and measuring the severity of the signs of depression (T2), where the participants were given the full user history and their systems had to automatically estimate the user's responses to a standard depression questionnaire.

Overall, the proposed tasks received 73 variants or runs from 12 teams. Although the effectiveness of the proposed solutions is still modest, the experiments suggest that evidence extracted from Social Media is valuable and automatic or semi-automatic screening tools could be designed to detect at-risk individuals. This promising result encourages us to further explore the creation of benchmarks for text-based screening of signs of risk.

Acknowledgements

We thank the support obtained from the Swiss National Science Foundation (SNSF) under the project "Early risk prediction on the Internet: an evaluation corpus", 2015.

This work was funded by FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación/ Project (RTI2018-093336-B-C21).

This work has also received financial support from the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2018/29) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System.

References

1. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An Inventory for Measuring Depression. JAMA Psychiatry **4**(6), 561–571 (06 1961)
2. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016. Evora, Portugal (2016)
3. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 346–360. Springer International Publishing, Cham (2017)
4. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2017. Dublin, Ireland (2017)
5. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2018. Avignon, France (2018)
6. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: Early Risk Prediction on the Internet. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 343–361. Springer International Publishing, Cham (2018)
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early risk prediction on the Internet. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 340–357. Springer International Publishing (2019)
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk at CLEF 2019: Early risk prediction on the Internet (extended overview). In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2019. Lugano, Switzerland (2019)
9. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: WSDM. pp. 495–503. ACM (2018)
10. Trotzek, M., Koitka, S., Friedrich, C.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering (04 2018)