Overview of eRisk at CLEF 2022: Early Risk Prediction on the Internet (Extended Overview)

Javier Parapar[1], Patricia Martín-Rodilla[1], David E. Losada[2] and Fabio Crestani[3]

[1]*Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña. Campus de Elviña s/n C.P 15071 A Coruña, Spain*

[2]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),Universidade de Santiago de Compostela. Rúa de Jenaro de la Fuente Domínguez, C.P 15782, Santiago de Compostela, Spain*

[3]*Faculty of Informatics, Universitá della Svizzera italiana (USI). Campus EST, Via alla Santa 1, 6900 Viganello, Switzerland*

Abstract

This paper provides an overview of eRisk 2022, the sixth edition of this lab, at the CLEF conference. Since its inception, the primary purpose of our lab has been to investigate topics concerning evaluation techniques, effectiveness metrics, and other processes connected to early risk detection on the internet. Early warning models can be employed in a range of contexts, including health and safety. This year, eRisk proposed three tasks. The first one was to discover early indicators of pathological gambling. The second task was to identify early signs of depression. The third required participant to automatically complete an eating disorders questionnaire (based on user writings on social media).

Keywords

early risk detection, pathological gambling, early detection of depression, eating disorders

1. Introduction

The major purpose of eRisk is to conduct research on evaluation methodologies, metrics, and other elements related to building research collections and identifying difficulties for early risk identification. Early detection technology can be useful in a variety of sectors, notably those involving safety and health. An automated system may issue early warnings when a person begins to exhibit indications of a mental illness, a sexual abuser begins engaging with a child, or a suspected criminal begins making antisocial threats on the Internet.

While our evaluation approaches (new research collections development strategies, creative evaluation measures, etc.) can be applied across different domains, eRisk has thus far concentrated on psychological difficulties (depression, self-harm, pathological gambling, and eating disorders).

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

javier.parapar@udc.es (J. Parapar); patricia.martin.rodilla@udc.es (P. Martín-Rodilla); david.losada@usc.es (D. E. Losada); fabio.crestani@usi.ch (F. Crestani)

https://www.dc.fi.udc.es/~parapar (J. Parapar); http://www.incipit.csic.es/gl/persoa/patricia-martin-rodilla (P. Martín-Rodilla); http://tec.citius.usc.es/ir/ (D. E. Losada); https://search.usi.ch/en/people/4f0dd874bbd63c00938825fae1843200/crestani-fabio (F. Crestani)

0000-0002-5997-8252 (J. Parapar); 0000-0002-1540-883X2 (P. Martín-Rodilla); 0000-0001-8823-7501 (D. E. Losada); 0000-0001-8672-0700 (F. Crestani)

In 2017, we conducted an exploratory task on the early detection of depression [1, 2]. The evaluation methods and test dataset described in [3] were the focus of this pilot task. In 2018, we continued detecting early signs of depression while also launching a new challenge of detecting early signs of anorexia [4, 5]. In 2019, we ran the continuation of the challenge on early identification of symptoms of anorexia, a challenge on early detection of signs of self-harm, and a third task aimed at estimating a user's responses to a depression questionnaire focused on her social media interactions [6, 7, 8]. In 2020, we continued with the early detection of self-harm, and the task on severity estimation of depression symptoms [9, 10, 11]. Finally, in the last edition in 2021, we presented two tasks on early detection (pathological gambling and self-harm) and one on the severity estimation of depression [12, 13, 14].

We've had the opportunity to compare a wide selection of solutions that use various technologies and concepts over the years (e.g. Natural Language Processing, Machine Learning, or Information Retrieval). We discovered that the link between psychological disorders and language use is complex, and that some contributing systems are not very effective. For example, the majority of participants had performance levels (e.g., F1) that were less than 70%. These figures indicate that further research into early prediction tasks is needed, and the solutions presented thus far have a lot of space for improvement.

In 2022, the lab had three campaign-style tasks [15]. The first task is the second edition of the pathological gambling domain. This task follows the same organisation as previous early detection challenges. The second task is also a continuation of the early detection of the depression challenge, whose last edition was in 2018. Finally, we provided a new task for the eating disorder severity estimation. Participants were required to analyse the user's posts and then estimate the user's answers to a standard eating disorder questionnaire. We describe these tasks in greater detail in the following sections of this overview article. We had 93 teams registered for the lab. We finally received results from 17 of them: 41 runs for Task 1, 62 runs for Task 2 and 12 for Task 3.

The lab had three campaign-style projects in 2022 [15].The first task is the second edition of the early alert problem on the pathological gambling domain. This challenge follows the same organisation as prior early detection challenges. The second one is likewise a continuation of the early identification of depression challenge, which had its most recent edition in 2018. Finally, we present a novel task for the eating disorder severity estimation. Participants have to analyze the user's posts and then estimate the user's responses to a typical eating disorder questionnaire. These tasks are described in greater detail in the next sections of this overview article. The lab had 93 teams registered. We eventually got responses from 17 of them: 41 runs for Task 1, 62 runs for Task 2, and 12 for Task 3.

2. Task 1: Early Detection of Signs of Pathological Gambling

This is a follow-up to Task 1 from 2021. The task was to create new models for detecting pathological gambling risk early on. Pathological gambling is also known as ludomania (ICD-10-CM code F63.0). It is usually referred as *gambling addiction* (an urge to gamble independently of its negative consequences). According to the World Health Organization [16], adult gambling addiction has prevalence rates ranging from 0.1 percent to 6.0 percent in 2017. As quickly as

Table 1
Task 1 (pathological gambling). Main statistics of the collection

	Train		Test	
	Gamblers	*Control*	*Gamblers*	*Control*
Num. subjects	164	2,184	81	1998
Num. submissions (posts & comments)	54,674	1,073,88	14,627	1,014,122
Avg num. of submissions per subject	333.37	491.70	180.58	507.56
Avg num. of days from first to last submission	≈ 560	≈ 662	≈ 489.7	≈ 664.9
Avg num. words per submission	30.64	20.08	30.4	22.2

feasible, the task required progressively digesting evidence and detecting early indicators of pathological gambling, also known as compulsive gambling or disordered gambling. Participating systems have to read and analyse Social Media posts in the order that users wrote them. As a result, systems that perform well in this task may be used to monitor user activities in blogs, social networks, and other types of online media in a sequential manner.

This task's test collection followed the same format as the collection specified in [3]. It is a collection of writings (posts or comments) from a set of Social Media users. The data source is also the same as in earlier eRisks (Reddit). There are two types of users: pathological gamblers and non-pathological gamblers, and the collection includes a series of posts for each (in chronological order). We put up a server that distributed user writings among the participating teams incrementally. The lab website[1] contains more information about the server. This was a training and test task. The teams got access to training data for the training stage, where we published the entire history of writings for training users. We identified which users had specifically said that they are pathological gamblers. As a result, the participants could tweak their systems using the training data. The training data for Task 1 in 2022 was made up of all Task 1 users from 2021.

Participants connected to our server and iteratively received user writings and sent responses during the test stage. At any time in the user chronology, any participant could pause and deliver an alert. After reading each user post, the teams had to decide whether to alert about the user (the system predicts the person would develop the risk) or not. Participants had to make this decision independently for each user in the test split. We regarded alerts as final (i.e. further decisions about this individual were ignored). In contrast, *no alerts* were regarded as provisional (i.e. the participants could later submit an alert about this user if they detected the appearance of signs of risk). To evaluate the systems, we used the correctness of the decisions and the number of user writes required to produce the decisions (see below).

We set up a REST service to help with testing. While waiting for responses, the server progressively disseminated user writings to each participant (no new user data was distributed to a specific participant until the service received all decisions for users and runs from that team in previous step). From January 17th, 2022 through April 22nd, 2022, the service was open for submissions. We used current methodologies that optimise the utilisation of assessors' time to produce the ground truth assessments [17, 18]. These methods enable the creation of test collections through the use of simulated pooling algorithms. The key statistics of the test

[1] https://early.irlab.org/server.html

collection utilised for T1 are reported in table 1 . The following sections go over evaluation methods.

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[2]), we computed $ERDE$, the early risk detection error used in previous editions of the lab. A full description of $ERDE$ can be found in [3]. 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 [19] 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 messages is exhausted. Thus, the performance metric should not depend on 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 [20]. 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

[2] computed with respect to the positive class.

do not want systems to detect these cases too late). Therefore, a first and intuitive measure of delay can be defined as follows[3]:

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

This measure of latency is calculated 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):

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

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

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (4)$$

Furthermore, Sadeque et al. proposed a measure, $F_{latency}$, which combines the effectiveness of the decision (estimated with the F measure) and the delay[4] in the decision. This is calculated by 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:

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

where p is a parameter that determines how quickly the penalty should increase. In [20], p was set such that the penalty equals 0.5 at the median number of posts of a user[5]. Observe that a decision right after the first writing has no penalty (i.e. $penalty(1) = 0$). Figure 1 plots how the latency penalty increases with the number of observed writings.
The system's overall speed factor is computed as:

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

where 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:

[3]Observe that Sadeque et al (see [20], 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.

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

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

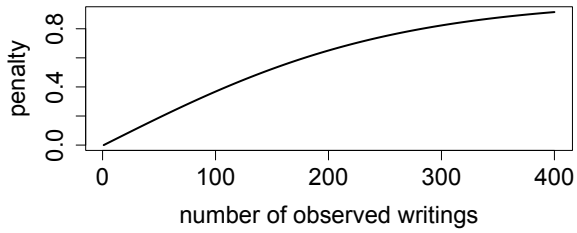


Figure 1: Latency penalty increases with the number of observed writings (k_u)

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

Since 2019 user's data were 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 explains a different type of evaluation that was employed in addition to the examination stated above. Following each data release (new user writing), participants were required to provide back the following information (for each user in the collection): i) a user decision (alert/no alert), which was utilised to compute the decision-based metrics outlined above, and ii) a score representing the user's level of risk (estimated from the evidence seen so far). We used these results to create a user ranking based on decreasing assessed risk. We have one ranking at each point for each participating system (i.e., ranking after one writing, ranking after two writings, etc.). This replicates a constant re-ranking strategy based on previous evidence. In practise, this ranking would be offered to an expert user who would make decisions (e.g. by inspecting the rankings).
Each ranking can be evaluated 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 2 shows the participating teams, the number of runs submitted and the approximate lapse of time from the first response to the last response. This time-lapse is indicative of the degree of

Table 2
Task 1 (pathological gambling): participating teams, number of runs, number of user writings processed by the team, and lapse of time taken for the entire process.

team	#runs	#user writings processed	lapse of time (from 1st to last response)
UNED-NLP	5	2001	17:58:48
SINAI	3	46	4 days 12:54:03
BioInfo_UAVR	5	1002	22:35:47
RELAI	5	109	7 days 15:27:25
BLUE	3	2001	3 days 13:15:25
BioNLP-UniBuc	5	3	00:37:33
UNSL	5	2001	1 day 21:53:51
NLPGroup-IISERB	5	1020	15 days 21:30:48
stezmo3	5	30	12:30:26

automation of each team's algorithms. A few of the submitted runs processed the entire thread of messages (2001), but many variants stopped earlier. Some of the teams were still submitting results at the deadline time. Three teams processed the thread of messages reasonably fast (around a day 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 extension 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, the best performing team was the NLPGroup-IISERB (run 4) but at the expense of a very low recall. In terms of $F1$, $ERDE_{50}$ and latency-weighted $F1$, the best performing run was submitted by the UNED NLD team. Their run (#4) also has a pretty high level of Recall (.938). Many teams achieved perfect Recall at the expense of very low Precision figures. In terms of $ERDE_5$, the best performing runs are SINAI #0 and #1 and BLUE #0. The majority of teams made quick decisions. Overall, these findings indicate that some systems achieved a relatively high level of effectiveness with only a few user submissions. Social and public health systems may use the best predictive algorithms to assist expert humans in detecting signs of pathological gambling as early as possible..

Table 4 presents the ranking-based results. Because some teams only processed a few dozens of user writings, we could only compute their user rankings for the initial number of processsed writings. For those participants providing ties in the scores for the users, we used the traditional *docid* criteria (subject name) for breaking the ties. Some runs (e.g., UNED-NLP #4, BLUE #0 and #1 and UNSL #0, #1 and #2) have very good levels of ranking-based shallow effectiveness over multiple points (after one writing, after 100 writings, and so forth). Regarding the 100 cut-off, the best performing teams after one writing for nDCG are UNED-NLP (#2) and BLUE (#0 and #1). In the other scenarios, both UNED-NLP and UNSL obtain very good results.

3. Task 2: Early Detection of Depression

This is a continuation of the tasks from 2017 and 2018. This task proposes early risk detection of depression in the same way as pathological gambling explained in Section 2. This task's

Table 3
Decision-based evaluation for Task 1

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNED-NLP	0	0.285	0.975	0.441	0.019	0.010	2.0	0.996	0.440
UNED-NLP	1	0.555	0.938	0.697	0.019	0.009	2.5	0.994	0.693
UNED-NLP	2	0.296	0.988	0.456	0.019	0.009	2.0	0.996	0.454
UNED-NLP	3	0.536	0.926	0.679	0.019	0.009	3.0	0.992	0.673
UNED-NLP	4	0.809	0.938	**0.869**	0.020	**0.008**	3.0	0.992	**0.862**
SINAI	0	0.425	0.765	0.546	**0.015**	0.011	**1.0**	**1.000**	0.546
SINAI	1	0.575	0.802	0.670	**0.015**	0.009	**1.0**	**1.000**	0.670
SINAI	2	0.908	0.728	0.808	0.016	0.011	**1.0**	**1.000**	0.808
BioInfo_UAVR	0	0.093	0.988	0.170	0.040	0.017	5.0	0.984	0.167
BioInfo_UAVR	1	0.067	**1.000**	0.126	0.047	0.024	5.0	0.984	0.124
BioInfo_UAVR	2	0.052	**1.000**	0.099	0.051	0.029	5.0	0.984	0.097
BioInfo_UAVR	3	0.050	**1.000**	0.095	0.052	0.030	5.0	0.984	0.094
BioInfo_UAVR	4	0.192	0.988	0.321	0.033	0.011	5.0	0.984	0.316
RELAI	0	0.000	0.000	0.000	0.039	0.039			
RELAI	1	0.000	0.000	0.000	0.039	0.039			
RELAI	2	0.052	0.963	0.099	0.036	0.029	**1.0**	**1.000**	0.099
RELAI	3	0.051	0.963	0.098	0.037	0.030	**1.0**	**1.000**	0.098
RELAI	4	0.000	0.000	0.000	0.039	0.039			
BLUE	0	0.260	0.975	0.410	**0.015**	0.009	**1.0**	**1.000**	0.410
BLUE	1	0.123	0.988	0.219	0.021	0.015	**1.0**	**1.000**	0.219
BLUE	2	0.052	**1.000**	0.099	0.037	0.028	**1.0**	**1.000**	0.099
BioNLP-UniBuc	0	0.039	**1.000**	0.075	0.038	0.037	**1.0**	**1.000**	0.075
BioNLP-UniBuc	1	0.039	**1.000**	0.076	0.038	0.037	**1.0**	**1.000**	0.076
BioNLP-UniBuc	2	0.040	**1.000**	0.077	0.037	0.036	**1.0**	**1.000**	0.077
BioNLP-UniBuc	3	0.046	**1.000**	0.087	0.033	0.032	**1.0**	**1.000**	0.087
BioNLP-UniBuc	4	0.046	**1.000**	0.089	0.032	0.031	**1.0**	**1.000**	0.089
UNSL	0	0.401	0.951	0.564	0.041	**0.008**	11.0	0.961	0.542
UNSL	1	0.461	0.938	0.618	0.041	**0.008**	11.0	0.961	0.594
UNSL	2	0.398	0.914	0.554	0.041	**0.008**	12.0	0.957	0.531
UNSL	3	0.365	0.864	0.513	0.017	0.009	3.0	0.992	0.509
UNSL	4	0.052	0.988	0.100	0.051	0.030	5.0	0.984	0.098
NLPGroup-IISERB	0	0.107	0.642	0.183	0.030	0.025	2.0	0.996	0.182
NLPGroup-IISERB	1	0.044	**1.000**	0.084	0.046	0.033	3.0	0.992	0.083
NLPGroup-IISERB	2	0.043	**1.000**	0.083	0.041	0.034	**1.0**	**1.000**	0.083
NLPGroup-IISERB	3	0.140	**1.000**	0.246	0.025	0.014	2.0	0.996	0.245
NLPGroup-IISERB	4	**1.000**	0.074	0.138	0.038	0.037	41.5	0.843	0.116
stezmo3	0	0.116	0.864	0.205	0.034	0.015	5.0	0.984	0.202
stezmo3	1	0.116	0.864	0.205	0.049	0.015	12.0	0.957	0.196
stezmo3	2	0.152	0.914	0.261	0.033	0.011	5.0	0.984	0.257
stezmo3	3	0.139	0.864	0.240	0.047	0.013	12.0	0.957	0.229
stezmo3	4	0.160	0.901	0.271	0.043	0.011	7.0	0.977	0.265

test collection followed the same format as the collection specified in [3]. The data source is also the same as in earlier eRisks. There are two types of users: depressed and non-depressed, and the collection offers a series of posts for each user (in chronological order). In contrast to earlier versions of the challenge, this is the first edition to use the REST service rather than the

Table 4
Ranking-based evaluation for Task 1

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
UNED-NLP	0	0.90	0.88	0.75	0.40	0.29	0.70	0.30	0.20	0.56	0.30	0.19	0.48
UNED-NLP	1	0.90	0.81	0.68	0.80	0.73	0.83	0.50	0.43	0.80	0.50	0.37	0.75
UNED-NLP	2	0.90	0.88	**0.76**	0.60	0.58	0.79	0.40	0.33	0.55	0.30	0.24	0.46
UNED-NLP	3	0.90	0.81	0.71	0.70	0.66	0.84	0.40	0.35	0.78	0.50	0.42	0.73
UNED-NLP	4	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.88	**1.00**	**1.00**	**0.95**	**1.00**	**1.00**	**0.95**
SINAI	0	0.10	0.19	0.56									
SINAI	1	0.70	0.65	0.62									
SINAI	2	**1.00**	**1.00**	0.70									
BioInfo_UAVR	0	0.00	0.00	0.03	0.80	0.87	0.33	0.00	0.00	0.00	0.10	0.10	0.03
BioInfo_UAVR	1	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioInfo_UAVR	2	0.00	0.00	0.03	0.40	0.30	0.29	0.00	0.00	0.02	0.10	0.19	0.05
BioInfo_UAVR	3	0.00	0.00	0.03	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.07	0.02
BioInfo_UAVR	4	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.03
RELAI	0	0.30	0.19	0.31	0.20	0.18	0.21						
RELAI	1	0.30	0.19	0.31	0.20	0.13	0.27						
RELAI	2	0.40	0.34	0.41	0.10	0.12	0.36						
RELAI	3	0.40	0.34	0.41	0.50	0.47	0.37						
RELAI	4	0.00	0.00	0.01	0.00	0.00	0.00						
BLUE	0	**1.00**	**1.00**	**0.76**	**1.00**	**1.00**	0.81	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.89
BLUE	1	**1.00**	**1.00**	**0.76**	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
BLUE	2	**1.00**	**1.00**	0.69	**1.00**	**1.00**	0.40	0.00	0.00	0.02	0.00	0.00	0.01
BioNLP-UniBuc	0	0.00	0.00	0.06									
BioNLP-UniBuc	1	0.00	0.00	0.02									
BioNLP-UniBuc	2	0.00	0.00	0.04									
BioNLP-UniBuc	3	0.10	0.19	0.07									
BioNLP-UniBuc	4	0.00	0.00	0.02									
UNSL	0	**1.00**	**1.00**	0.68	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	0.93	**1.00**	**1.00**	**0.95**
UNSL	1	**1.00**	**1.00**	0.70	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	0.92	**1.00**	**1.00**	0.93
UNSL	2	0.90	0.90	0.66	**1.00**	**1.00**	0.77	0.90	0.92	0.78	0.90	0.90	0.77
UNSL	3	**1.00**	**1.00**	0.69	0.60	0.58	0.72	0.80	0.81	0.77	0.80	0.81	0.78
UNSL	4	0.10	0.07	0.32	0.10	0.07	0.32	0.20	0.13	0.33	0.30	0.22	0.37
NLPGroup-IISERB	0	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.03
NLPGroup-IISERB	1	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.00	0.03
NLPGroup-IISERB	2	0.00	0.00	0.15	0.00	0.00	0.11	0.20	0.13	0.12	0.00	0.00	0.08
NLPGroup-IISERB	3	0.00	0.00	0.01	0.10	0.06	0.10	0.10	0.07	0.12	0.10	0.07	0.12
NLPGroup-IISERB	4	0.20	0.38	0.15	0.00	0.00	0.06	0.00	0.00	0.07	0.00	0.00	0.07
stezmo3	0	0.10	0.06	0.26									
stezmo3	1	0.10	0.06	0.26									
stezmo3	2	0.50	0.58	0.61									
stezmo3	3	0.50	0.58	0.61									
stezmo3	4	0.50	0.58	0.61									

chuck-based release. The lab website[6]contains more information about the server.
This was a training and test task. The test phase was conducted in the same manner as Task 1 (see Section 2). The teams got access to training data for the training stage, where we published the entire history of writings for training users. We highlighted those who had expressly stated that they suffer from depression. As a result, the participants could tweak their systems using

[6]https://early.irlab.org/server.html

Table 5
Task 2 (Depression). Main statistics of test collection

	Test	
	Depressed	*Control*
Num. subjects	98	1,302
Num. submissions (posts & comments)	35,332	687,228
Avg num. of submissions per subject	360.53	527,82
Avg num. of days from first to last submission	$\approx$ 628.2	$\approx$ 661.7
Avg num. words per submission	27.4	23.5

Table 6
Task 2 (depression): 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 processed	lapse of time (from 1st to last response)
CYUT	5	2000	7 days 12:02:44
LauSAn	5	2000	2 days 06:44:17
BLUE	3	2000	2 days 17:16:05
BioInfo_UAVR	5	503	09:38:26
TUA1	5	2000	16:28:49
NLPGroup-IISERB	5	632	11 days 20:35:11
RELAI	5	169	7 days 02:27:10
UNED-MED	5	1318	5 days 13:18:24
Sunday-Rocker2	5	682	4 days 03:54:25
SCIR2	5	2000	1 day 04:52:02
UNSL	5	2000	1 day 09:35:12
E8-IJS	5	2000	3 days 02:36:32
NITK-NLP2	4	6	01:52:57

the training data. The training data for Task 2 in 2022 was made up of users from the 2017 and 2018 editions.
Again, we followed existing methods to build the assessments using simulated pooling strategies, which optimize the use of assessors time [17, 18]. Table 5 reports the main statistics of the test collections used for T2. The same decision and ranking based measures as discussed in sections 2.1 and 2.2 were used for this task.

3.1. Task 2: 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. Most of the submitted runs processed the entire thread of messages (about 2000), but few stopped earlier or were not able to process the users' history in time. Only one team was able to process the entire set of writings in less than a day.

Table 7: Decision-based evaluation for Task 2

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
CYUT	0	0.165	0.918	0.280	0.053	0.032	3.0	0.992	0.277
CYUT	1	0.162	0.898	0.274	0.053	0.032	3.0	0.992	0.272
CYUT	2	0.106	0.867	0.189	0.056	0.047	**1.0**	**1.000**	0.189
CYUT	3	0.149	0.878	0.255	0.075	0.040	7.0	0.977	0.249
CYUT	4	0.142	0.918	0.245	0.082	0.041	8.0	0.973	0.239
LauSAn	0	0.137	0.827	0.235	0.041	0.038	**1.0**	**1.000**	0.235
LauSAn	1	0.165	0.888	0.279	0.053	0.040	2.0	0.996	0.278
LauSAn	2	0.174	0.867	0.290	0.056	0.031	4.0	0.988	0.287
LauSAn	3	0.420	0.643	0.508	0.059	0.041	6.0	0.981	0.498
LauSAn	4	0.201	0.724	0.315	**0.039**	0.033	**1.0**	**1.000**	0.315
BLUE	0	0.395	0.898	0.548	0.047	0.027	5.0	0.984	0.540
BLUE	1	0.213	0.939	0.347	0.054	0.033	4.5	0.986	0.342
BLUE	2	0.106	**1.000**	0.192	0.074	0.048	4.0	0.988	0.190
BioInfo_UAVR	0	0.222	0.949	0.360	0.071	0.031	9.0	0.969	0.349
BioInfo_UAVR	1	0.091	0.969	0.166	0.101	0.054	8.0	0.973	0.162
BioInfo_UAVR	2	0.171	0.969	0.291	0.083	0.035	11.0	0.961	0.279
BioInfo_UAVR	3	0.090	0.990	0.166	0.101	0.052	6.0	0.981	0.162
BioInfo_UAVR	4	0.378	0.857	0.525	0.069	0.031	16.0	0.942	0.494
TUA1	0	0.155	0.806	0.260	0.055	0.037	3.0	0.992	0.258
TUA1	1	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1	2	0.155	0.806	0.260	0.055	0.037	3.0	0.992	0.258
TUA1	3	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1	4	0.159	0.959	0.272	0.052	0.036	3.0	0.992	0.270
NLPGroup-IISERB	0	0.682	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**
NLPGroup-IISERB	1	0.385	0.857	0.532	0.062	0.032	18.0	0.934	0.496
NLPGroup-IISERB	2	0.662	0.459	0.542	0.069	0.058	62.0	0.766	0.416
NLPGroup-IISERB	3	0.653	0.500	0.566	0.067	0.046	26.0	0.903	0.511
NLPGroup-IISERB	4	0.000	0.000	0.000	0.070	0.070			
RELAI	0	0.085	0.847	0.155	0.114	0.092	51.0	0.807	0.125
RELAI	1	0.085	0.847	0.155	0.114	0.091	51.0	0.807	0.125
RELAI	2	0.000	0.000	0.000	0.070	0.070			
RELAI	3	0.000	0.000	0.000	0.070	0.070			
RELAI	4	0.000	0.000	0.000	0.070	0.070			
UNED-MED	0	0.119	0.969	0.212	0.091	0.056	18.0	0.934	0.198
UNED-MED	1	0.139	0.980	0.244	0.079	0.046	13.0	0.953	0.233
UNED-MED	2	0.122	0.939	0.215	0.086	0.057	15.0	0.945	0.204
UNED-MED	3	0.131	0.949	0.231	0.084	0.051	15.0	0.945	0.218
UNED-MED	4	0.084	0.163	0.111	0.079	0.078	251.0	0.252	0.028
Sunday-Rocker2	0	0.091	**1.000**	0.167	0.080	0.053	4.0	0.988	0.165
Sunday-Rocker2	1	0.355	0.786	0.489	0.068	0.041	27.0	0.899	0.439
Sunday-Rocker2	2	0.092	0.388	0.149	0.088	0.083	117.5	0.575	0.085
Sunday-Rocker2	3	0.283	0.816	0.420	0.071	0.045	37.5	0.859	0.361
Sunday-Rocker2	4	0.108	**1.000**	0.195	0.082	0.047	6.0	0.981	0.191
SCIR2	0	0.396	0.837	0.538	0.076	0.076	150.0	0.477	0.256
SCIR2	1	0.336	0.878	0.486	0.078	0.078	150.0	0.477	0.232
SCIR2	2	0.235	0.908	0.373	0.051	0.046	3.0	0.992	0.370
SCIR2	3	0.316	0.847	0.460	0.079	**0.026**	44.0	0.834	0.383
SCIR2	4	0.274	0.847	0.414	0.045	0.031	3.0	0.992	0.411
UNSL	0	0.161	0.918	0.274	0.079	0.042	14.5	0.947	0.260
UNSL	1	0.310	0.786	0.445	0.078	0.037	12.0	0.957	0.426
UNSL	2	0.400	0.755	0.523	0.045	**0.026**	3.0	0.992	0.519
UNSL	3	0.144	0.929	0.249	0.055	0.035	3.0	0.992	0.247

Table 7: Decision-based evaluation for Task 2 (Continuation)

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL	4	0.080	0.918	0.146	0.099	0.074	5.0	0.984	0.144
E8-IJS	0	**0.684**	0.133	0.222	0.061	0.061	**1.0**	**1.000**	0.222
E8-IJS	1	0.242	0.959	0.387	0.068	0.036	20.5	0.924	0.357
E8-IJS	2	0.000	0.000	0.000	0.070	0.070			
E8-IJS	3	0.000	0.000	0.000	0.070	0.070			
E8-IJS	4	0.000	0.000	0.000	0.070	0.070			
NITK-NLP2	0	0.138	0.796	0.235	0.047	0.039	2.0	0.996	0.234
NITK-NLP2	1	0.135	0.806	0.231	0.047	0.039	2.0	0.996	0.230
NITK-NLP2	2	0.132	0.786	0.225	0.050	0.040	2.0	0.996	0.225
NITK-NLP2	3	0.149	0.724	0.248	0.049	0.039	2.0	0.996	0.247

Table 7 reports the decision-based performance achieved by the participating teams. In terms of Precision, E8-IJS run #0 obtains the highest values but at the expenses of low Recall. Similarly, Sunday-Rocker systems #0 and #4 obtain and BLUE #2 perfect Recall but with low Precision values. When considering the Precision-Recall trade-off, NLPGroup-IISERB #0 is the best performance being the only run over 0.7 (highest $F1$). Regarding latency-penalized metrics, UNSL #2 and SCIR2 #3 obtained the best $ERDE_{50}$ and LauSAn #4 the best $ERDE_5$ error value. It is again NLPGroup-IISERB #04, the one achieving the best latency-weighted $F1$. This run seems to be quite balanced overall.

Table 8: Ranking-based evaluation for Task 2

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
CYUT	0	0.50	0.49	0.37	0.50	0.52	0.54	0.60	0.59	0.58	0.70	0.72	0.61
CYUT	1	0.70	0.77	0.37	0.60	0.72	0.58	0.60	0.72	0.61	0.70	0.80	0.62
CYUT	2	0.00	0.00	0.16	0.10	0.07	0.25	0.10	0.19	0.31	0.10	0.12	0.29
CYUT	3	0.10	0.07	0.12	0.70	0.70	0.57	0.70	0.72	0.59	**0.80**	0.74	0.60
CYUT	4	0.10	0.06	0.12	0.60	0.68	0.55	0.60	0.69	0.59	**0.80**	0.84	0.61
LauSAn	0	0.60	0.72	0.43	0.30	0.41	0.13	0.20	0.31	0.12	0.10	0.19	0.11
LauSAn	1	0.60	0.66	0.43	0.40	0.33	0.30	0.50	0.50	0.17	0.20	0.15	0.08
LauSAn	2	0.60	0.66	0.43	0.40	0.33	0.29	0.50	0.50	0.18	0.20	0.15	0.13
LauSAn	3	0.60	0.66	0.43	0.40	0.33	0.27	0.50	0.50	0.22	0.20	0.15	0.14
LauSAn	4	0.40	0.38	0.34	0.50	0.49	0.41	0.40	0.27	0.21	0.20	0.22	0.14
BLUE	0	**0.80**	**0.88**	**0.54**	0.60	0.56	0.59	0.80	0.81	0.66	**0.80**	0.80	0.68
BLUE	1	**0.80**	**0.88**	**0.54**	0.70	0.64	**0.67**	0.80	0.84	**0.74**	**0.80**	**0.86**	**0.72**
BLUE	2	**0.80**	0.75	0.46	0.40	0.40	0.30	0.30	0.35	0.20	0.30	0.38	0.16
BioInfo_UAVR	0	0.00	0.00	0.04	0.20	0.15	0.15	0.00	0.00	0.09			
BioInfo_UAVR	1	0.00	0.00	0.02	0.20	0.25	0.14	0.20	0.12	0.07			
BioInfo_UAVR	2	0.20	0.13	0.06	0.60	0.60	0.36	0.70	0.78	0.32			
BioInfo_UAVR	3	0.10	0.08	0.08	0.20	0.26	0.14	0.20	0.17	0.08			
BioInfo_UAVR	4	0.10	0.07	0.05	0.00	0.00	0.04	0.00	0.00	0.05			

Table 8: Ranking-based evaluation for Task 2 (Continuation)

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	P@10	$NDCG$@10	$NDCG$@100	P@10	$NDCG$@10	$NDCG$@100	P@10	$NDCG$@10	$NDCG$@100	P@10	$NDCG$@10	$NDCG$@100
TUA1	0	**0.80**	**0.88**	0.44	0.60	0.72	0.52	0.60	0.67	0.52	0.70	0.80	0.57
TUA1	1	0.70	0.77	0.44	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1	2	**0.80**	**0.88**	0.44	0.60	0.72	0.52	0.60	0.67	0.52	0.70	0.80	0.57
TUA1	3	0.60	0.69	0.43	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1	4	0.50	0.37	0.35	0.00	0.00	0.36	0.00	0.00	0.36	0.20	0.12	0.31
NLPGroup-IISERB	0	0.00	0.00	0.02	**0.90**	0.92	0.30	**0.90**	**0.92**	0.33			
NLPGroup-IISERB	1	0.30	0.32	0.13	**0.90**	0.81	0.27	0.80	0.84	0.33			
NLPGroup-IISERB	2	0.70	0.79	0.24	0.00	0.00	0.00	0.00	0.00	0.00			
NLPGroup-IISERB	3	0.00	0.00	0.06	0.10	0.19	0.06	0.00	0.00	0.02			
NLPGroup-IISERB	4	0.00	0.00	0.04	**0.90**	0.93	0.66	**0.90**	**0.92**	0.69			
RELAI	0	0.00	0.00	0.07	0.10	0.06	0.20						
RELAI	1	0.00	0.00	0.07	0.20	0.25	0.20						
RELAI	2	0.10	0.12	0.09	0.00	0.00	0.16						
RELAI	3	0.10	0.12	0.09	0.50	0.52	0.31						
RELAI	4	0.10	0.12	0.07	0.00	0.00	0.00						
UNED-MED	0	0.70	0.69	0.27	0.80	0.84	0.63	0.60	0.66	0.60	0.50	0.46	0.56
UNED-MED	1	0.50	0.44	0.26	0.70	0.76	0.50	0.60	0.64	0.47	0.80	0.74	0.50
UNED-MED	2	0.70	0.68	0.28	0.50	0.51	0.59	0.80	0.71	0.61	0.50	0.44	0.62
UNED-MED	3	**0.80**	0.82	0.29	0.60	0.44	0.31	0.80	0.73	0.36	0.40	0.51	0.30
UNED-MED	4	0.00	0.00	0.06	0.00	0.00	0.05	0.00	0.00	0.04	0.10	0.19	0.09
Sunday-Rocker2	0	0.40	0.47	0.39	0.40	0.44	0.29	0.50	0.46	0.24			
Sunday-Rocker2	1	0.70	0.81	0.39	**0.90**	**0.93**	0.66	**0.90**	0.88	0.65			
Sunday-Rocker2	2	0.10	0.07	0.23	0.00	0.00	0.11	0.30	0.31	0.17			
Sunday-Rocker2	3	**0.80**	**0.88**	0.41	0.50	0.50	0.23	0.60	0.69	0.34			
Sunday-Rocker2	4	0.30	0.28	0.31	0.30	0.37	0.25	0.40	0.30	0.18			
SCIR2	0	0.10	0.07	0.08	0.00	0.00	0.06	0.00	0.00	0.06	0.10	0.12	0.06
SCIR2	1	0.00	0.00	0.05	0.10	0.07	0.07	0.00	0.00	0.04	0.00	0.00	0.05
SCIR2	2	0.00	0.00	0.06	0.00	0.00	0.05	0.20	0.13	0.07	0.00	0.00	0.06
SCIR2	3	0.10	0.06	0.05	0.00	0.00	0.04	0.00	0.00	0.06	0.00	0.00	0.02
SCIR2	4	0.10	0.19	0.09	0.10	0.07	0.05	0.10	0.10	0.07	0.10	0.06	0.05
UNSL	0	0.60	0.40	0.36	0.20	0.13	0.46	0.30	0.28	0.43	0.60	0.72	0.45
UNSL	1	**0.80**	**0.88**	0.46	0.60	0.73	0.64	0.60	0.73	0.66	0.60	0.71	0.66
UNSL	2	0.70	0.68	0.50	0.50	0.39	0.55	0.70	0.73	0.61	0.70	0.73	0.61
UNSL	3	0.10	0.06	0.15	0.40	0.27	0.43	0.30	0.21	0.42	0.30	0.21	0.42
UNSL	4	0.10	0.12	0.05	0.00	0.00	0.03	0.20	0.19	0.07	0.00	0.00	0.04
E8-IJS	0	0.00	0.00	0.06	0.10	0.07	0.05	0.10	0.12	0.08	0.00	0.00	0.03
E8-IJS	1	0.40	0.58	0.19	0.40	0.41	0.15	0.20	0.15	0.09	0.30	0.38	0.15
E8-IJS	2	0.00	0.00	0.05	0.00	0.00	0.07	0.00	0.00	0.05	0.10	0.19	0.07
E8-IJS	3	0.00	0.00	0.02	0.10	0.10	0.08	0.10	0.06	0.02	0.00	0.00	0.05
E8-IJS	4	0.00	0.00	0.07	0.10	0.10	0.08	0.20	0.31	0.11	0.10	0.06	0.04
NITK-NLP2	0	0.40	0.28	0.15									
NITK-NLP2	1	0.00	0.00	0.01									
NITK-NLP2	2	0.00	0.00	0.02									
NITK-NLP2	3	0.00	0.00	0.02									

Table 8 presents the ranking-based results. Contrary to task 1, no run obtained perfect figures for any of the scenarios. This is worth noting, given that for task 2, there are more positive subjects. Overall, systems #0 and #1 from the BLUE team seem to be the most consistent under the different number of writings among the best-performing ones. Other systems, such as those from NLPGroup-IISERB, show an erratic behaviour going so low as Precision 0 when only one

writing was processed but obtaining the best results for the same metrics after 100.

4. Task 3: Measuring the severity of the signs of Eating Disorders

The challenge consists on estimating the severity of various symptoms linked with an eating disorder diagnosis. In order to accomplish this, the participants worked from a thread of user posts. Participants were given a whole history of Social Media posts and comments for each user, and they had to evaluate the individual's replies to a typical eating disorder questionnaire (based on the evidence revealed in the history of posts/comments).
The questionnaire is based on the Eating Disorder Examination Questionnaire (EDE-Q)[7], which is a 28-item self-reported questionnaire derived from the semi-structured interview Eating Disorder Examination (EDE)[8][21]. We only used questions 1 through 12 and 19 through 28. This test is intended to measure the extent and severity of multiple eating disorder symptoms. It incorporates four subscales (Restraint, Eating Concern, Shape Concern, and Weight Concern) as well as an overall score. Table 9 has a list of questions.

Table 9: Eating Disorder Examination Questionarie

```
Instructions:

The following questions are concerned with the past four weeks (28 days) only. Please read each
question carefully. Please answer all the questions. Thank you..

1. Have you been deliberately trying to limit the amount of food you eat to influence your shape
or weight (whether or not you have succeeded)  0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

2. Have you gone for long periods of time (8 waking hours or more) without eating anything at all
in order to influence your shape or weight?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

3. Have you tried to exclude from your diet any foods that you like in order to influence your
shape or weight (whether or not you have succeeded)?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY
```

[7] https://www.corc.uk.net/media/1273/ede-q_quesionnaire.pdf
[8] https://www.corc.uk.net/media/1951/ede_170d.pdf

Table 9: Eating Disorder Examination Questionarie (continued)

4. Have you tried to follow definite rules regarding your eating (for example, a calorie limit) in order to influence your shape or weight (whether or not you have succeeded)?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

5. Have you had a definite desire to have an empty stomach with the aim of influencing your shape or weight?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

6. Have you had a definite desire to have a totally flat stomach?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

7. Has thinking about food, eating or calories made it very difficult to concentrate on things you are interested in (for example, working, following a conversation, or reading)?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

8. Has thinking about shape or weight made it very difficult to concentrate on things you are interested in (for example, working, following a conversation, or reading)?
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

9. Have you had a definite fear of losing control over eating
 0. NO DAYS
 1. 1-5 DAYS
 2. 6-12 DAYS
 3. 13-15 DAYS
 4. 16-22 DAYS
 5. 23-27 DAYS
 6. EVERY DAY

10. Have you had a definite fear that you might gain weight?

Table 9: Eating Disorder Examination Questionarie (continued)

0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

11. Have you felt fat?
0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

12. Have you had a strong desire to lose weight?
0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

19. Over the past 28 days, on how many days have you eaten in secret (ie, furtively)? 0... Do not count episodes of binge eating.
0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

20. On what proportion of the times that you have eaten have you felt guilty (felt that you've done wrong) because of its effect on your shape or weight? 0... Do not count episodes of binge eating.
0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

21. Over the past 28 days, how concerned have you been about other people seeing you eat? 0... Do not count episodes of binge eating
0. NO DAYS
1. 1-5 DAYS
2. 6-12 DAYS
3. 13-15 DAYS
4. 16-22 DAYS
5. 23-27 DAYS
6. EVERY DAY

22. Has your weight influenced how you think about (judge) yourself as a person?
0. NOT AT ALL (0)
1. SLIGHTY (1)

Table 9: Eating Disorder Examination Questionarie (continued)

2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

23. Has your shape influenced how you think about (judge) yourself as a person?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

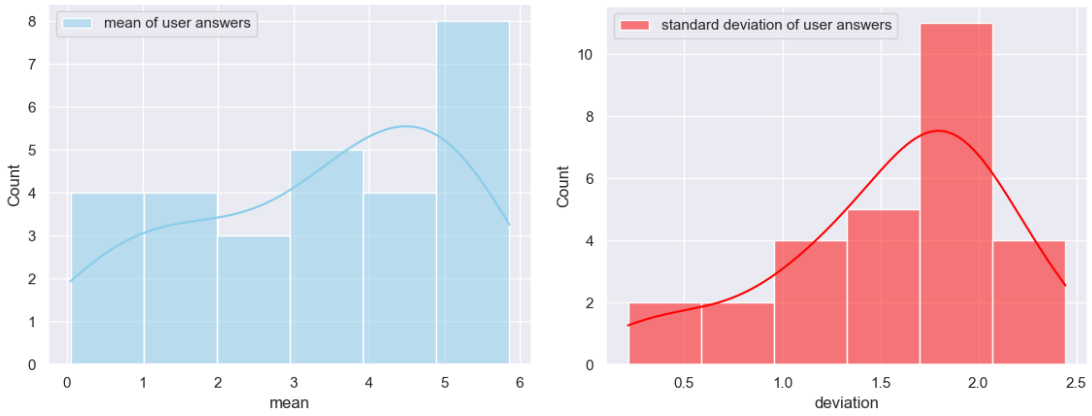
24. How much would it have upset you if you had been asked to weigh yourself once a week (no more, or less, often) for the next four weeks?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

25. How dissatisfied have you been with your weight?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

26. How dissatisfied have you been with your shape?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

27. How uncomfortable have you felt seeing your body (for example, seeing your shape in the mirror, in a shop window reflection, while undressing or taking a bath or shower)?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)

28. How uncomfortable have you felt about others seeing your shape or figure (for example, in communal changing rooms, when swimming, or wearing tight clothes)?
0. NOT AT ALL (0)
1. SLIGHTY (1)
2. SLIGHTY (2)
3. MODERATELY (3)
4. MODERATELY (4)



(a) Mean. **(b)** Standard deviation.

Figure 2: User answers mean and standard deviation.

Table 9: Eating Disorder Examination Questionarie (continued)

5. MARKEDLY (5)
6. MARKEDLY (6)

The goal of this task is to study the feasibility of automatically evaluating the severity of multiple eating disorder symptoms. Based on the user's writing history, the algorithms must estimate the user's response to each specific question. We collected surveys filled out by Social Media users as well as their writing history (we extracted each history of writings right after the user provided us with the filled questionnaire). The quality of the responses produced by the participating systems was evaluated using user-completed questionnaires (ground truth). This was a test task only. The participants received no training data.
The participants were given a dataset with 28 individuals (each user's writing history was provided) and asked to create a file with the following structure:

```
username1 answer1 answer2 ... answer22
username2 answer1 answer2 ... answer22
```

Each line has the username and 22 values. These values correspond with the responses to the questions above (the possible values are 0,1,2,3,4,5,6). The next figure illustrates the mean and standard deviation of user answers considering all the questions.

4.1. Task 3: Evaluation Metrics

Evaluation is based on the following effectiveness metrics:

- **Mean Zero-One Error (**$MZOE$**)** between the questionnaire filled by the real user and the questionnaire filled by the system (i.e. fraction of incorrect predictions).

$$MZOE(f, Q) = \frac{|\{q_i \in Q : R(q_i) \neq f(q_i)\}|}{|Q|} \tag{8}$$

where f denotes the classification done by an automatic system, Q is the set of questions of each questionnaire, q_i is the i-th question, $R(q_i)$ is the real user's answer for the i-th question and $f(q_i)$ is the predicted answer of the system for the i-th question. Each user produces a single $MZOE$ score and the reported $MZOE$ is the average over all $MZOE$ values (mean $MZOE$ over all users).

- **Mean Absolute Error (MAE)** between the questionnaire filled by the real user and the questionnaire filled by the system (i.e. average deviation of the predicted response from the true response).

$$MAE(f, Q) = \frac{\sum_{q_i \in Q} |R(q_i) - f(q_i)|}{|Q|} \tag{9}$$

Again, each user produces a single MAE score and the reported MAE is the average over all MAE values (mean MAE over all users).

- **Macroaveraged Mean Absolute Error (MAE_{macro})** between the questionnaire filled by the real user and the questionnaire filled by the system (see [22]).

$$MAE_{macro}(f, Q) = \frac{1}{7} \sum_{j=0}^{6} \frac{\sum_{q_i \in Q_j} |R(q_i) - f(q_i)|}{|Q_j|} \tag{10}$$

where Q_j represents the set of questions whose true answer is j (note that j goes from 0 to 6 because those are the possible answers to each question). Again, each user produces a single MAE_{macro} score and the reported MAE_{macro} is the average over all MAE_{macro} values (mean MAE_{macro} over all users).

The following measures are based on aggregated scores obtained from the questionnaires. Further details about the EDE-Q instruments can be found elsewhere (e.g. see the scoring section of the questionnaire[9]).

- **Restraint Subscale (RS)**: Given a questionnaire, its restraint score is obtained as the mean response to the first five questions. This measure computes the RMSE between the restraint ED score obtained from the questionnaire filled by the real user and the restraint ED score obtained from the questionnaire filled by the system.

Each user u_i is associated with a real subscale ED score (referred to as $R_{RS}(u_i)$) and an estimated subscale ED score (referred to as $f_{RS}(u_i)$). This metric computes the RMSE between the real and an estimated subscale ED scores as follows:

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} (R_{RS}(u_i) - f_{RS}(u_i))^2}{|U|}} \tag{11}$$

where U is the user set.

[9] https://www.corc.uk.net/media/1951/ede_170d.pdf

- **Eating Concern Subscale (ECS)**: Given a questionnaire, its eating concern score is obtained as the mean response to the following questions (7, 9, 19, 21, 20). This metric computes the RMSE (equation 12) between the eating concern ED score obtained from the questionnaire filled by the real user and the eating concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f,U) = \sqrt{\frac{\sum_{u_i \in U}(R_{ECS}(u_i) - f_{ECS}(u_i))^2}{|U|}} \tag{12}$$

- **Shape Concern Subscale (SCS)**: Given a questionnaire, its shape concern score is obtained as the mean response to the following questions (6, 8, 23, 10, 26, 27, 28, 11). This metric computes the RMSE (equation 13) between the shape concern ED score obtained from the questionnaire filled by the real user and the shape concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f,U) = \sqrt{\frac{\sum_{u_i \in U}(R_{SCS}(u_i) - f_{SCS}(u_i))^2}{|U|}} \tag{13}$$

- **Weight Concern Subscale (WCS)**: Given a questionnaire, its weight concern score is obtained as the mean response to the following questions (22, 24, 8, 25, 12). This metric computes the RMSE (equation 14) between the weight concern ED score obtained from the questionnaire filled by the real user and the weight concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f,U) = \sqrt{\frac{\sum_{u_i \in U}(R_{WCS}(u_i) - f_{WCS}(u_i))^2}{|U|}} \tag{14}$$

- **Global ED (GED)**: To obtain an overall or 'global' score, the four subscales scores are summed and the resulting total divided by the number of subscales (i.e. four) [21]. This metric computes the RMSE between the real and an estimated global ED scores as follows:

$$RMSE(f,U) = \sqrt{\frac{\sum_{u_i \in U}(R_{GED}(u_i) - f_{GED}(u_i))^2}{|U|}} \tag{15}$$

4.2. Task 3: Results

Table 10 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 6s, which consist of sending the same response (0 or 6) for all the questions, and average, which is the performance achieved by a method that, for each question, sends as a response the answer that is the closest to the mean of the responses sent by all participants (e.g. if the mean response provided by the participants equals 3.7 then this average approach would submit a 4). Table 11 reports the names of the runs, and the **team** and corresponding **run ID**.

Table 10
Task 3 Results. Participating teams and runs with corresponding scores for the metrics.

team	run ID	MZOE	MAE	MAE_{macro}	GED	RS	ECS	SCS	WCS
NLPGroup-IISERB	1	0.92	2.58	2.09	2.04	2.16	1.89	2.74	2.33
NLPGroup-IISERB	2	0.92	**2.18**	**1.76**	**1.74**	**2.00**	**1.73**	**2.03**	**1.92**
NLPGroup-IISERB	3	0.93	2.60	2.10	2.04	2.13	1.90	2.74	2.35
NLPGroup-IISERB	4	0.81	3.36	2.96	3.68	3.69	3.18	4.28	3.82
RELAI	1	0.82	3.31	2.91	3.59	3.65	3.05	4.19	3.74
RELAI	2	0.82	3.30	2.89	3.56	3.65	3.03	4.17	3.71
RELAI	3	0.83	3.15	2.70	3.26	3.04	2.72	4.04	3.61
RELAI	4	0.82	3.32	2.91	3.59	3.66	3.05	4.19	3.74
RELAI	5	0.82	3.19	2.74	3.34	3.15	2.80	4.08	3.64
SINAI	1	0.85	2.65	2.29	2.63	3.29	2.35	2.98	2.40
SINAI	2	0.87	2.60	2.23	2.42	3.01	2.21	2.85	2.31
SINAI	3	0.86	2.62	2.22	2.54	3.15	2.32	2.93	2.36
all 0		0.81	3.36	2.96	3.68	3.69	3.18	4.28	3.82
all 6		**0.67**	2.64	3.04	3.25	3.52	3.72	2.81	3.28
average		0.88	2.72	2.22	2.69	2.76	2.20	3.35	2.85

Table 11
Task 3. Teams, names of runs and run IDs

team	run ID	name
NLPGroup-IISERB	1	IISERB_bert
NLPGroup-IISERB	2	IISERB_jl
NLPGroup-IISERB	3	IISERB_pretrained_bert
NLPGroup-IISERB	4	IISERB_sp
RELAI	1	RELAI_A-W2V-ED
RELAI	2	RELAI_B-Glove200-ED
RELAI	3	RELAI_B-Glove200-ED-AllFeatures
RELAI	4	RELAI_B-W2V-ED
RELAI	5	RELAI_B-W2V-ED-AllFeartures
SINAI	1	SINAI_0.4
SINAI	2	SINAI_0.35
SINAI	3	SINAI_0.375

4.2.1. Distribution of MAE and MZOE

Figure 3 plots for each system the distribution of MAE across the set of test users, while Figure 4 summarizes the MZOE values obtained by each system. Each box plot gives an idea about the distribution of the effectiveness metric over the available users.

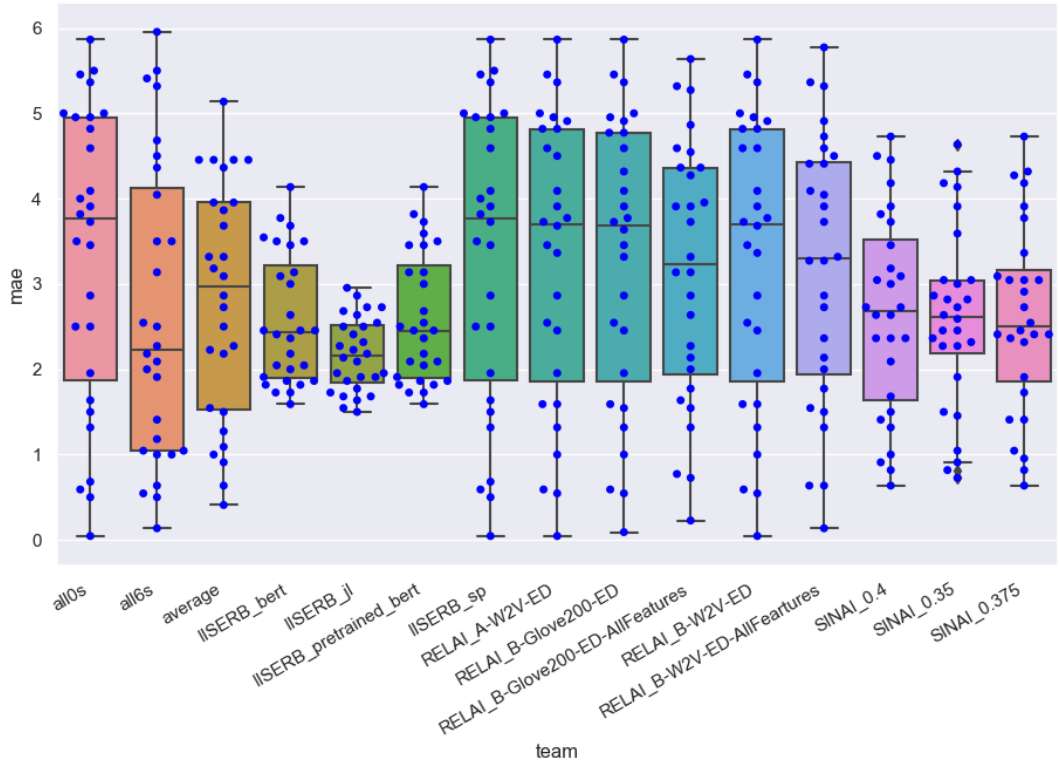


Figure 3: MAE evaluation per user.

5. Participating Teams

Table 12 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 2021 working notes proceedings.

NLPGroup-IISERB [23]. The NLP-IISERB Lab participated in the three tasks proposed as part of eRisk CLEF this year. Regarding Task 1 and Task 2, the team performed five runs using different text mining frameworks in which AB, LR, RF and SVM classifiers were tested, as well as BERT, Bio-BERT, RoBERTa and Longformer models from the HuggingFace library. All of them with a variety of engineering features and techniques. The results achieved for Task 1 and Task 2 present successful numbers on precision, recall and F1. NLPGroup-IISERB run #4 achieves the best precision score (1.0) among the precision scores of all 41 submissions for task 1 of the eRisk2022 challenge. The team observed from the empirical analysis that the classical BOW model performs better than all the deep learning-based models on the given data except the longformer model. The Longformer model performed as good as the BOW model for Task1, but we could not explore its performance for task2 owing to time limitations. Regarding task 3, NLP-IISERB were one of the top results in the competition, presenting four runs. NLPGroup-IISERB run #2, a combination of cosine similarity and BERT model fine-tuned

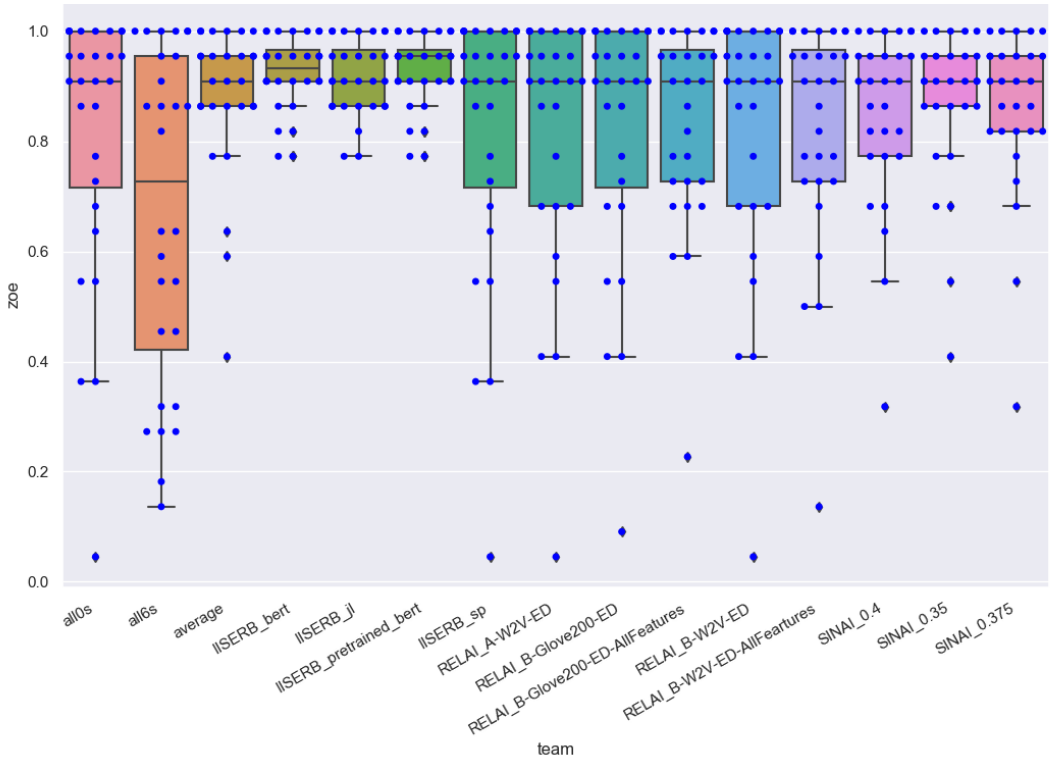


Figure 4: MZOE evaluation per user.

on anorexia dataset from eRisk 2018 shared task 2 performed the best among all the other runs for task 3 in terms of all the evaluation metrics except MZOE metric. The proposed models performed well in terms of GED score, indicating that they reasonably identify eating disorders and their side effects.

RELAI [24]. Their working notes present the similarity-based approaches proposed by the RELAI (Université du Québec and McMaster University) team to Task 3 of eRisk 2022. The proposed methods rely on feature sets dedicated to each item in the questionnaire. The feature sets are compared to the written production of users based on pre-trained word vectors. The developed models try to measure the severity of the signs of ED in an unsupervised manner. Thus, the philosophy of the approach is based on dedicating sets of characteristics to each element of the questionnaire. They compare pre-trained vectors with those generated for each item and try to measure their severity in an unsupervised way. Two similarity-based models (each with four variations) and 22 feature sets designed using expert knowledge were developed. Two kinds of pre-trained word vectors were used as word representations. The first is 300- dimensional word2vec, word vectors trained on publicly available textual content such as Wikipedia and UMBC WebBase corpus using the Continuous Bag Of Words (CBOW) model. The second is GloVe word vectors trained on two billion Twitter posts (tweets).

BioInfo_UAVR [25]. The University of Aveiro participated only in tasks 1 and 2 this year. Their approach was all centred on finding the best feature engineering technique, that is, finding

Table 12
eRisk 2022 participants

team	Task 1 #runs	Task 2 #runs	Task 3
NLPGroup-IISERB	5	5	4
RELAI	5	5	5
BioInfo_UAVR	5	5	
BLUE	3	3	
UNSL	5	5	
SINAI	3		3
UNED-NLP	5		
BioNLP-UniBuc	5		
stezmo3	5		
CYUT		5	
LauSAn		5	
TUA1		5	
UNED-MED		5	
Sunday-Rocker2		5	
E8-IJS		5	
NITK-NLP2		5	
SCIR2		5	

the most useful textual features. They tested Bag-of-Words with tf-idf, GoVe word embedding, and contextualized language model. These were tested alone and with sentiment analysis. The results show that these techniques achieved very high recall and the expenses of very low precision. In addition, the use of sentiment analysis did not improve performance.

BLUE [26]. The BLUE team represents a joint collaboration between the University of Bucharest (Romania) and Universitat Politécnica de Valencia (Spain). This team participated in the two early detection challenges (T1 and T2) and employed a transformer-based architecture for user-level classification. More specifically, the interaction between users' posts was analysed, and some technological elements were oriented to mitigate noise in the dataset. Within this process, the system learns to ignore uninformative posts. This team also made important efforts to facilitate interpretability.

UNSL [27] The UNSL team includes researchers affiliated with the Universidad Nacional de San Luis (Argentina), the Consejo Nacional de Investigaciones Científicas y Técnica de Argentina, Instituto de Matemática Aplicada San Luis (Argentina), and the IDIAP Research Institute from Switzerland. The group participated in tasks 1 and 2 on early risk detection. Their proposal is built upon their models from previous editions. In particular, they proposed two new policies for their EarlyModel. The historic stop policy uses a rolling window over the last decision probabilities of the model. If the probabilities on the window exceed a threshold, then the model emits an alert. And the "learned decision tree stop policy" that authors only used on task 2 that learned a decision tree over the depression dataset from 2018.

SINAI [28] The SINAI team is affiliated with the Universidad de Jaen, Spain. They participated in tasks 1 and 3. For task 1, they devised an approach based on regression over RoBERTa sentence embeddings using different features: volumetry, lexical diversity, text complexity

and emotions. For task 3, they used the last 28 days of the users' history and used embedding similarities between questions from the EDE-Q and user writings. For the day-based questions, they counted the number of days with similarities higher than a threshold and defined similarity ranges for the scale-based questions.

UNED-NLP [29] The UNED-NLP is a joint collaboration between Universidad Nacional de Educación a Distancia (UNED) and the Instituto Mixto de Investigación de la Escuela Nacional de Sanidad (IMIENS), Spain. The team participated in task 1. The group explored the use of approximate nearest neighbours (ANN) over dense representations of the users' posts based on the Universal Sentence Encoder. Post-level labels were re-computed from the user-level annotations in the training data.

BioNLP-UniBuc [30] The team from the University of Bucharest participated in Task 1. After processing the provided XML files, the resulting training dataset had an unbalanced number of examples for each of the two classes, so they used a stratified 5-fold cross-validation to try to reduce the class imbalance in the train/validation splits. For feature extraction, the authors used the Bag-of-Words and tf-idf models with additional properties for extracting relevant features such as removing the rare words or frequent words, constructing 2-grams and 3-grams. The team performs three runs combining classification methods with deep learning models. When using deep learning models, the best outcome was obtained by a model containing a hidden dimension of 128 of the linear projection, an embedding size of 64 tokens, and four attention heads.

stezmo3 [31] The ZHAW team is affiliated to Zurich University of Applied Sciences and their experiments focused on reproducing the solutions proposed by UNSL for eRisk 2021 and, additionally, incorporating some innovations related to the use of Glove to support feature extraction. This team participated in eRisk 2022 T1 task, on early identification of pathological gambling. All variants proposed are based on Glove features, extracted from user posts and the classification of user posts was done with SVMs or XGBoost.

CYUT [32] THE CYUT team belongs to the Chaoyang University of Technology, Taiwan. They participated in the early risk detection of depression task (task 2). Their runs exploit RoBERTa massive pre-trained model to address the problem with out-of-the-box and improved representations. In their best run (#2), they use the output vector of the last layer hidden for linear classification as post representation.

LauSAn [33] The University of Zurich took part only in the early detection of depression, and they focused expressively on the time-sensitivity of the task. Their approach is based on two simple strategies of optimising standard text classification models to the early detection of depression: concatenating the posts in different ways during and inference training (to optimise training), and continuously changing the decision threshold at inference time (to optimise the results of time-sensitive metrics). An ablation study confirmed that both strategies were effective. In fact, the team achieved among the best results in all time-sensitive evaluation metrics.

TUA1 [34] The University of Tokushima participated only in the detection of early risks of depression. They proposed a novel and very interesting approach called TAM: Time-Aware-Affective Memories. A TAM network maintains the memory of a user's "affective state" which gets updated as a new user's posting becomes available. All of this is fed to a Transformer decoder which then predicts the user's risk of depression. A study of the latency penalty

complements the approach to see how this could be effectively used to reduce the ERDE metric score. Their results show that the approach is very efficient and particularly good for the very early-stage detection of depression.

UNED-MED [35] The team from the Spanish National University of Distance Education (UNED) propose two rather standard approaches. The first is based on the use of feature-driven classifiers employing features based on textual data, like terms tf-idf, first-person pronoun use, sentiment analysis and depression terminology. The second is on a Deep Learning classifier with pre-trained Embeddings. The main innovation is to enlarge the training data (to make it more balanced) with data extracted from the same source of eRisk data. Yet the results show only modest performance. An explanation suggested by the teams is that the depression detection task was more challenging this year. This might be possible, but it is something the organisers will need to crosscheck across all participants.

Sunday-Rocker2 [36] The team from the University of Bucharest applied a variety of techniques for addressing the T2 challenge, from using tf-idf and linguistic features extracted from Reddit writings to using the MixUP technique (a new approach for sentence classification to augment the data through interpolation). The main interest is in the novelty of the approaches taken and the way that the features were selected, with some strongly linguistic-based features, such as self-preoccupation (based on the use of the occurrence of first-number pronouns) and similar features. The best results were obtained with a Voting Classifier with hard voting applied on textual features and with an SVM used on both textual features and numerical features as well.

E8-IJS [37] This Slovenian team includes researchers from Jozef Stefan Institute (Ljubljana) and the Faculty of Computer and Information Sciences. Their participation in eRisk 2022 focused on the task of early identification of depression cases. To that end, these participants utilized a classical supervised learning approach (Logistic Regression), and the main goal of their experiments was to compare different input representations to the logistic classifier. This included experiments with i) tf-idf representations that were reduced to a latent space via Latent Semantic Analysis and ii) Sentence Bert representations of the input data. Their classification methods work at the post level.

NITK-NLP2 [38] This team is affiliated to the National Institute of Technology Surathkal at Karnataka (India). These participants employed BERT-based and DeBERTa-based models to classify user's posts. Their solutions included data augmentation methods to deal with imbalance and the team focused on comparing the relative performance of the two transformer-based methods.

SCIR2 This team is affiliated to the Harbin Institute of Technology in Heilongjiang, China. Their experiments focused on the use of RoBERTa models, together with techniques aimed at reducing the number of user posts that are employed for analysis. This team participated in the early detection of depression task (T2).

6. Conlusions

The purpose of this paper was to present an overview of eRisk 2022. This lab's sixth version focused on two sorts of tasks. On the one hand, two tasks focused on early identification

of pathological gambling and depression (Tasks 1 and 2, respectively), in which participants were given sequential access to the user's social media posts and were required to issue alerts regarding at-risk persons. On the other hand, one task (Task 3) was assigned to measuring the severity of the indicators of eating disorders, in which participants were provided the whole user history and their systems were required to automatically predict the user's replies to a standard depression questionnaire. The proposed tasks received 115 runs from 17 different teams. Although the efficacy of the proposed solutions is currently limited, the experimental results demonstrate that evidence retrieved from social media is valuable, and automatic or semi-automatic screening methods to discover at-risk persons might be developed. These findings push us to investigate the establishment of text-based risk indicator screening benchmarks.

Acknowledgements

This work was supported by projects PLEC2021-007662 (MCIN/AEI/10.13039/ 501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU) and RTI2018-093336-B-C21, RTI2018-093336-B-C22 (Ministerio de Ciencia e Innvovación & ERDF). The first and second authors thank the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G/01, GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT of the University of A Coruña as a Research Center of the Galician University System. The third author also thanks the financial support supplied by 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] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations, in: G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2017, pp. 346–360.

[2] D. E. Losada, F. Crestani, J. Parapar, 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.

[3] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016, Evora, Portugal, 2016.

[4] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early Risk Prediction on the Internet, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan,

L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 343–361.
[5] D. E. Losada, F. Crestani, J. Parapar, 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] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019: Early risk prediction on the Internet, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, 2019, pp. 340–357.
[7] D. E. Losada, F. Crestani, J. Parapar, 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.
[8] D. E. Losada, F. Crestani, J. Parapar, Early detection of risks on the internet: An exploratory campaign, in: Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part II, 2019, pp. 259–266.
[9] D. E. Losada, F. Crestani, J. Parapar, 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, 2020, pp. 272–287.
[10] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at CLEF 2020: Early risk prediction on the internet (extended overview), in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, 2020.
[11] D. E. Losada, F. Crestani, J. Parapar, erisk 2020: Self-harm and depression challenges, in: Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, 2020, pp. 557–563.
[12] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, 2021, pp. 324–344.
[13] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, 2021, pp. 864–887.
[14] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2021: Pathological gambling, self-harm and depression challenges, in: Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, 2021, pp. 650–656.
[15] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2022: Pathological gambling, depression, and eating disorder challenges, in: Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, 2022, pp. 436–442.
[16] M. Abbott, The epidemiology and impact of gambling disorder and other gambling-related harm, in: WHO Forum on alcohol, drugs and addictive behaviours, Geneva, Switzerland,

2017.
[17] D. Otero, J. Parapar, Á. Barreiro, Beaver: Efficiently building test collections for novel tasks, in: Proceedings of the First Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020, 2020.
[18] D. Otero, J. Parapar, Á. Barreiro, The wisdom of the rankers: a cost-effective method for building pooled test collections without participant systems, in: SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021, 2021, pp. 672–680.
[19] M. Trotzek, S. Koitka, C. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Transactions on Knowledge and Data Engineering (2018).
[20] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: WSDM, ACM, 2018, pp. 495–503.
[21] C. G. Fairburn, Z. Cooper, M. O'Connor, Eating disorder examination Edition 17.0D (April, 2014).
[22] S. Baccianella, A. Esuli, F. Sebastiani, Evaluation measures for ordinal regression, 2009, pp. 283–287. doi:10.1109/ISDA.2009.230.
[23] H. Srivastava, L. Ns, S. S, T. Basu, Nlp-iiserb@erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[24] S. H. H. Saravani, D. Maupomé, F. Rancourt, T. Soulas, L. Normand, S. Besharati, S. M. Anaelle Normand, M.-J. Meurs, Measuring the severity of the signs of eating disorders using similarity-based models, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[25] R. Ferreira, A. Trifan, J. L. Oliveira, Early risk detection of mental illnesses using various types of textual features, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[26] A.-M. Bucur, A. Cosma, L. P. Dinu, P. Rosso, An end-to-end set transformer for user-level classification of depression and gambling disorder, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[27] J. M. Loyola, H. Thompson, S. Burdisso, M. Errecalde., Decision policies with history for early classification, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[28] A. M. Mármol-Romero, S. M. Jiménez-Zafra, F. M. P. del Arco, M. D. Molina-González, M.-T. Martín-Valdivia, A. Montejo-Ráez., Sinai at erisk@clef 2022, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[29] H. Fabregat, A. Duque, L. Araujo, J. Martinez-Romo, Uned-nlp at erisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[30] T.-A. Dumitrascu, A. M. Enescu, Clef erisk 2022: Detecting early signs of pathological

gambling using ml and dl models with dataset chunking, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[31] S. Stalder, E. Zankov, Zhaw at erisk 2022: Predicting signs of pathological gambling - glove for snowy days, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[32] S.-H. Wu, Z.-J. Qiu, Cyut at erisk 2022: Early detection of depression based-on concatenating representation of multiple hidden layers of roberta model, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[33] A. Säuberli, S. Cho, L. Stahlhut, Lausan at erisk 2022: Simply and effectively optimizing text classification for early detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[34] K. Xin, D. Rongyu, Y. Haitao, Tua1 at erisk 2022: Exploring affective memories for early detection of depression, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[35] E. Campillo-Ageitos, J. Martinez-Romo, L. Araujo, Uned-med at erisk 2022: depression detection with tf-idf, linguistic features and embeddings, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[36] R.-A. Gînga, A.-A. Manea, B.-M. Dobre, Sunday rockers at erisk 2022: Early detection of depression, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[37] I. Tavchioski, B. Škrlj, S. Pollak, B. Koloski, Early detection of depression with linear models using hand-crafted and contextual features, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.
[38] S. Devaguptam, T. Kogatam, N. Kotian, A. K. M., Early detection of depression using bert and deberta, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologina, Italy, September 5-8, 2022.