Overview of eRisk at CLEF 2021: 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 gives an outline of eRisk 2021, the CLEF conference's fifth edition of this lab. The main goal of eRisk is to explore issues of evaluation methodology, effectiveness metrics and other processes related to early risk detection. Early alerting models may be used in a variety of situations, including those involving health and safety. This edition of eRisk had three tasks. The first task focused on early detecting signs of pathological gambling. The second challenge was to spot early signs of self-harm. The third one required participants to fill out a depression questionnaire automatically based on user writings on social media.

Keywords

early risk, erisk, pathological gambling, self-harm, depression,

1. Introduction

The primary goal of eRisk is to investigate topics such as evaluation methodologies, metrics, and other factors relevant to developing research collections and identifying problems for early risk identification. Early detection technologies have the potential to be useful in a variety of fields, especially those related to safety and health. Early alerts may be issued, for example, when a person begins to exhibit symptoms of a psychotic illness, when a sexual abuser begins interacting with an infant, or when a suspected criminal begins publishing antisocial threats on the Internet.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

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)

While the evaluation methodology (strategies for developing new research sets, innovative evaluation metrics, etc.) can be extended across various domains, eRisk has so far concentrated on psychological issues (essentially, depression, self-harm and eating disorders). We conducted an exploratory task on the early diagnosis of depression in 2017 [1, 2]. This pilot task was focused on the evaluation methods and test dataset described in [3]. In 2018, we continued the task on early identification of symptoms of depression while also launching a new task on early detection of 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]. Finally, in 2020, we continued with the early detection of self-harm and the task on severity estimation of depression symptoms [9, 10, 11].
Over the years, we've been able to compare a variety of solutions that use diverse technologies and models (e.g. Natural Language Processing, Machine Learning, or Information Retrieval). We discovered that the interplay between psychological disorders and language use is challenging and that the effectiveness of most contributing systems is low. For example, most participants had performance levels (e.g., in terms of F1) that were less than 70%. This suggests that this kind of early prediction tasks requires additional investigation, and the solutions offered so far have a lot of space for improvement.
In 2021, the lab had three campaign-style tasks [12]. The first task explores a new domain: pathological gambling. We designed this new task in the same fashion as previous early detection challenges. The second task is a continuation of the early detection of the self-harm task. Finally, we provided the third edition of the depression severity estimation task, where participants were required to analyse the user's posts and then estimate the user's answers to a standard depression questionnaire. These tasks are described in greater detail in the next sections of this overview article. We had 76 teams registered for the lab. We finally received results from 18 of them: 26 runs for Task 1, 55 runs for Task 2 and 36 for Task 3.

2. Task 1: Early Detection of Pathological Gambling

This was a new task in 2021. The challenge was to conduct a study on early risk detection of pathological gambling. Pathological gambling (ICD-10-CM code F63.0) is also called ludomania and usually referred to as *gambling addiction* (it is an urge to gamble independently of its negative consequences). According to the World Health Organization [13], in 2017, adult gambling addiction had prevalence rates ranged from 0.1% to 6.0%. The task entailed sequentially processing evidence and detecting early signs of pathological gambling, also known as compulsive gambling or disordered gambling, as soon as possible. The task is primarily concerned with evaluating Text Mining solutions and focuses on texts written in Social Media. Participating systems had to read and process the posts in the order in which they were created on Social Media. As a result, systems that effectively perform this task could be used to sequentially monitor user interactions in blogs, social networks, and other types of online media.
The test collection for this task had the same format as the collection described in [3]. 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, pathological

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

	Test	
	Pathological Gamblers	*Control*
Num. subjects	164	2,184
Num. submissions (posts & comments)	54,674	1,073,883
Avg num. of submissions per subject	333.37	491.70
Avg num. of days from first to last submission	≈ 560	≈ 662
Avg num. words per submission	30.64	20.08

gamblers and non-pathological gamblers, and, for each user, the collection contains a sequence of writings (in chronological order). We set up a server that iteratively gave user writings to the participating teams. More information about the server can be found at the lab website[1].
This was an "*only test*" task. No training data was provided to the participants. The test stage consisted of participants connecting to our server and iteratively receiving user writings and sending responses. At any point in the user chronology, each participant could stop and issue an alert. After reading each user post, the teams had to choose between: i) alerting about the user (the system predicts the user will develop the risk) or ii) not alerting about the user. Alerts were regarded 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 about this user if they detected the appearance of signs of risk). This choice had to be made for each user in the test split. The accuracy of the decisions and the number of user writings required to make the decisions were used to evaluate the systems (see below). To support the testing stage, we deployed a REST service. The server iteratively distributed user writings to each participant while waiting for their responses (no new user data was distributed to a specific participant until the service received a decision from that team). The service was open for submissions from February 1st, 2021, until April 23rd 2021.
In order to build the ground truth assessments, we followed existing approaches that optimize the use of assessors time [14, 15]. These methods allow to build test collections using simulated pooling strategies. Table 1 reports the main statistics of the test collection used for T1. Evaluation measures are discussed in the next sections.

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.

[1] https://early.irlab.org/server.html
[2] computed with respect to the positive class.

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 [16] 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 [17]. 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[3]:

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

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):

[3]Observe that Sadeque et al (see [17], 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.

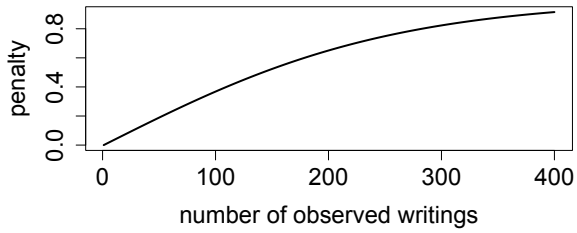

Figure 1: Latency penalty increases with the number of observed writings (k_u)

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

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

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

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 [17], 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:

[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.

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

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 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 2
Participating teams in Task 1: 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)
RELAI	5	1231	9 days 05:42:11
UPV-Symanto	5	801	18:42:54
UNSL	5	2000	5 days 01:23:26
BLUE	5	1828	1 days 23:43:28
CeDRI	2	271	1 days 05:44:10
EFE	4	2000	3 days 03:02:22

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 3
Decision-based evaluation for Task 1

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL	0	0.326	0.957	0.487	0.079	0.023	11	0.961	0.468
UNSL	1	0.137	0.982	0.241	0.060	0.035	4	0.988	0.238
UNSL	2	**0.586**	0.939	**0.721**	0.073	**0.020**	11	0.961	**0.693**
UNSL	3	0.084	0.963	0.155	0.066	0.060	1	**1**	0.155
UNSL	4	0.086	0.933	0.157	0.067	0.060	1	**1**	0.157
RELAI	0	0.138	0.988	0.243	**0.048**	0.036	1	**1**	0.243
RELAI	1	0.108	**1**	0.194	0.057	0.045	1	**1**	0.194
RELAI	2	0.071	**1**	0.132	0.067	0.064	1	**1**	0.132
RELAI	3	0.071	**1**	0.132	0.066	0.064	1	**1**	0.132
RELAI	4	0.070	**1**	0.131	0.066	0.065	1	**1**	0.131
BLUE	0	0.107	0.994	0.193	0.067	0.046	2	0.996	0.192
BLUE	1	0.157	0.988	0.271	0.054	0.036	2	0.996	0.270
BLUE	2	0.121	0.994	0.215	0.065	0.045	2	0.996	0.215
BLUE	3	0.095	**1**	0.174	0.071	0.051	2	0.996	0.173
BLUE	4	0.110	0.994	0.198	0.068	0.048	2	0.996	0.197
UPV-Symanto	0	0.042	0.415	0.077	0.088	0.087	1	**1**	0.077
UPV-Symanto	1	0.040	0.457	0.074	0.097	0.091	1	**1**	0.074
UPV-Symanto	2	0.030	0.238	0.053	0.093	0.091	1	**1**	0.053
UPV-Symanto	3	0.035	0.409	0.064	0.098	0.097	1	**1**	0.064
UPV-Symanto	4	0.028	0.256	0.051	0.098	0.095	1	**1**	0.051
CeDRI	0	0.076	**1**	0.142	0.079	0.060	2	0.996	0.141
CeDRI	1	0.070	**1**	0.131	0.066	0.065	1	**1**	0.131
EFE	0	0.251	0.640	0.361	0.079	0.037	16	0.942	0.340
EFE	1	0.296	0.537	0.382	0.076	0.043	31	0.884	0.337
EFE	2	0.233	0.750	0.356	0.082	0.033	11	0.961	0.342
EFE	3	0.292	0.549	0.381	0.076	0.044	31	0.884	0.337

automation of each team's algorithms. A few of the submitted runs processed the entire thread of messages (2000), but many variants opted for stopping earlier. Three teams processed the thread of messages in a reasonably fast way (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 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, $F1$, $ERDE_{50}$ and latency-weighted $F1$, the best performing run was submitted by the UNSL team. This run (#2) also has a quite high level of Recall (.939). Many teams achieved perfect Recall at the expense of very low Precision figures. In terms of $ERDE_5$, 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
UNSL	0	**1**	**1**	0.81	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
UNSL	1	**1**	**1**	0.79	0.8	0.73	0.87	0.8	0.69	0.86	0.8	0.62	0.84
UNSL	2	**1**	**1**	**0.85**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
UNSL	3	0.9	0.92	0.74	**1**	**1**	0.76	**1**	**1**	0.72	**1**	**1**	0.72
UNSL	4	**1**	**1**	0.69	0	0	0.25	0	0	0.11	0	0	0.13
RELAI	0	0.9	0.92	0.73	**1**	**1**	0.93	**1**	**1**	0.92	**1**	**1**	0.91
RELAI	1	**1**	**1**	0.72	**1**	**1**	0.91	**1**	**1**	0.91	**1**	**1**	0.91
RELAI	2	0.8	0.81	0.49	0.5	0.43	0.32	0.5	0.55	0.42	0.5	0.55	0.41
RELAI	3	0.8	0.88	0.61	0.6	0.68	0.49	0.7	0.77	0.55	0.8	0.85	0.55
RELAI	4	0.6	0.63	0.45	0	0	0.04	0	0	0.03	0	0	0.07
BLUE	0	0.9	0.88	0.61	0.8	0.73	0.57	0.9	0.93	0.64	0.7	0.78	0.60
BLUE	1	**1**	**1**	0.61	0.8	0.82	0.53	**1**	**1**	0.56	**1**	**1**	0.56
BLUE	2	0.6	0.70	0.73	0.8	0.87	0.76	0.8	0.88	0.75	0.9	0.90	0.76
BLUE	3	0.6	0.65	0.60	0.8	0.87	0.61	0.7	0.71	0.60	0.7	0.67	0.60
BLUE	4	0.9	0.81	0.73	**1**	**1**	0.77	**1**	**1**	0.76	**1**	**1**	0.78
UPV-Symanto	0	0	0	0	0	0	0	0	0	0			
UPV-Symanto	1	0	0	0	0	0	0	0	0	0			
UPV-Symanto	2	0	0	0	0	0	0	0	0	0			
UPV-Symanto	3	0	0	0	0	0	0	0	0	0			
UPV-Symanto	4	0	0	0	0	0	0	0	0	0			
CeDRI	0	0.9	0.93	0.64	0.7	0.63	0.40						
CeDRI	1	0	0	0.02	0	0	0.03						
EFE	0	0.2	0.29	0.28	0.6	0.64	0.52	0.6	0.62	0.53	0.6	0.62	0.52
EFE	1	0.2	0.29	0.28	0.6	0.64	0.52	0.6	0.62	0.53	0.6	0.62	0.52
EFE	2	0.5	0.45	0.40	0.6	0.56	0.50	0.6	0.57	0.54	0.6	0.57	0.52
EFE	3	0.5	0.45	0.40	0.6	0.56	0.50	0.6	0.57	0.54	0.6	0.57	0.52

best performing run is RELAI #0. This run, however, shows poor performance in terms of classification accuracy. 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 dozen 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.

Some runs (e.g., UNSL runs #0 #1 #2, RELAI #2) have the same levels of ranking-based shallow effectiveness over multiple points (after one writing, after 100 writings, and so forth). However, for the 100 cut-off, only UNSL #2 obtains the highest NDCG after one writing. This run is

Table 5
Task 2 (self-harm). Main statistics of test collection

	Train		Test	
	Self-Harm	*Control*	*Self-Harm*	*Control*
Num. subjects	145	618	152	1296
Num. submissions (posts & comments)	18,618	254,642	51,104	688,823
Avg num. of submissions per subject	128.4	412.0	336.2	531.5
Avg num. of days from first to last submission	$\approx$ 312	$\approx$ 461	$\approx$ 346	$\approx$ 510
Avg num. words per submission	22.4	15.2	26.03	20.74

consistently the best performing one in terms of ranking for every cut-off, metric and number of writings. The UPV-Symanto team seems to have some bug on their model as it consistently yielded zero performance.
In summary, UNSL #2 is overall the best performing run in ranking and decision-based evaluation.

3. Task 2: Early Detection of Self-Harm

This is a continuation of 2019 task 2 and 2020 task 1. This task proposes the early risk detection of self-harm in the very same way as described for pathological gambling in Section 2. The test collection for this task also had the same format as the collection described in [3]. The source of data is also the same used for previous eRisks. Here 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). We set up a server that iteratively gave user writings to the participating teams. More information about the server can be found at the lab website[6].
This was a train and a test task. The test phase followed the same procedure as Task 1 (see Section 2). For the training stage, the teams had access to training data where we released the whole history of writings for training users. We indicated what users had explicitly mentioned that they had done self-harm. The participants could therefore tune their systems with the training data. In 2021, the training data for Task 2 was composed of all 2019's Task 2 users and 2020's Task 1 test users.
Again, we followed existing methods to build the assessments using simulated pooling strategies, which optimize the use of assessors time [14, 15]. Table 5 reports the main statistics of the train and 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. The 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 (about 2000), but many variants opted for stopping earlier or were

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

Table 6
Participating teams in Task 2: 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)
NLP-UNED	5	472	07:08:37
AvocadoToast	3	379	10 days 13:20:37
Birmingham	5	11	2 days 08:01:32
NuFAST	3	6	17:07:57
NaCTeM	5	1999	5 days 20:22:04
EFE	4	1999	1 days 15:17:18
BioInfo@UAVR	2	91	1 days 02:21:30
NUS-IDS	5	46	3 days 08:11:46
RELAI	5	1561	11 days 00:49:27
CeDRI	3	369	1 days 09:51:27
BLUE	5	156	1 days 04:57:23
UPV-Symanto	5	538	11:56:33
UNSL	5	1999	3 days 17:36:10

not able to process the users' history in time. Only one team was able to process the entire set of writings in a reasonable amount of time (around a day or so for processing the entire history of user messages). The remaining teams took several days to complete the process. Some teams required more than a week. Again, this suggests that they used some form of offline processing.
Table 7 reports the decision-based performance achieved by the participating teams. In terms of Precision, Birmingham run #2 obtains the highest values but at the expenses of low Recall. Similarly, CEDRI systems #1 and #2 obtain perfect Recall but with low Precision values. When considering the Precision-Recall trade-off, UNSL #4 is the best performance being the only run over 0.6 (highest $F1$). Regarding latency-penalized metrics, UPV-Symanto #1 obtains the best $ERDE_5$ and UNSL #0 the best $ERDE_5$ error value. It is again UNSL #4, the one achieving the best latency-weighted $F1$. This run seems to be quite balanced overall. When comparing the best values with the ones from last year, the best values for Precision and F1 are lower than those reported in 2020. This year the amount of released training data more than doubled, but the availability of a larger training set was apparently no beneficial for the 2021 participants. Therefore, these results seem to suggest the need of models that better exploit existing information.
Table 8 presents the ranking-based results. Some runs perform equally for some of the ranking-based effectiveness over different cut-off values (e.g., UNSL runs #0 #3 #4 after one writing or NLP-UNED#4, BLUE #2 or UPV-Symanto #0 and #3 after 100 writings). After 500 and 1000 writings, RELAI #1 obtains the best values for shallow cut-offs. UNSL #4 obtains the highest NDCG and Precision at the 10 cut-off after one writing and very good values under the other situations. This seems to point out that this effective run keeps the same good overall behaviour as in the case of the decision-based evaluation.

Table 7
Decision-based evaluation for Task 2

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
NLP-UNED	0	0.442	0.75	0.556	0.080	0.042	6	0.981	0.545
NLP-UNED	1	0.442	0.796	0.568	0.091	0.041	11	0.961	0.546
NLP-UNED	2	0.422	0.73	0.535	0.088	0.047	7	0.977	0.522
NLP-UNED	3	0.419	0.77	0.543	0.093	0.047	10	0.965	0.524
NLP-UNED	4	0.453	0.816	0.582	0.088	0.040	9	0.969	0.564
AvocadoToast	0	0.214	0.757	0.334	0.111	0.069	11	0.961	0.321
AvocadoToast	1	0.245	0.401	0.304	0.078	0.076	**1**	**1**	0.304
AvocadoToast	2	0.215	0.757	0.335	0.111	0.069	11	0.961	0.322
Birmingham	0	0.584	0.526	0.554	0.068	0.054	2	0.996	0.551
Birmingham	1	0.644	0.309	0.418	0.097	0.074	8	0.973	0.406
Birmingham	2	**0.757**	0.349	0.477	0.085	0.070	4	0.988	0.472
Birmingham	3	0.629	0.434	0.514	0.084	0.062	5	0.984	0.506
Birmingham	4	0	0	0	0.105	0.105			
NuFAST	0	0.124	0.283	0.172	0.101	0.097	**1**	**1**	0.172
NuFAST	1	0.124	0.283	0.172	0.101	0.097	**1**	**1**	0.172
NuFAST	2	0.124	0.283	0.172	0.101	0.097	**1**	**1**	0.172
NaCTeM	0	0.108	0.882	0.193	0.185	0.184	1999	0.0	0.0
NaCTeM	1	0.108	0.882	0.193	0.185	0.184	1999	0.0	0.0
NaCTeM	2	0.108	0.882	0.193	0.185	0.184	1999	0.0	0.0
NaCTeM	3	0.108	0.882	0.193	0.184	0.184	1999	0.0	0.0
NaCTeM	4	0.108	0.882	0.193	0.184	0.184	1999	0.0	0.0
EFE	0	0.381	0.717	0.498	0.118	0.050	17	0.938	0.467
EFE	1	0.434	0.605	0.505	0.114	0.063	32	0.880	0.445
EFE	2	0.366	0.796	0.501	0.120	0.043	12	0.957	0.48
EFE	3	0.422	0.605	0.497	0.114	0.063	32	0.88	0.437
BioInfo@UAVR	0	0.233	0.862	0.367	0.136	0.050	22	0.918	0.337
BioInfo@UAVR	1	0.274	0.789	0.407	0.128	0.047	22	0.918	0.374
NUS-IDS	0	0.133	0.987	0.234	0.108	0.073	3	0.992	0.232
NUS-IDS	1	0.131	0.98	0.232	0.116	0.073	4	0.988	0.229
NUS-IDS	2	0.134	0.993	0.236	0.117	0.072	4	0.988	0.233
NUS-IDS	3	0.128	0.987	0.227	0.106	0.075	3	0.992	0.225
NUS-IDS	4	0.135	0.987	0.237	0.104	0.071	3	0.992	0.235
RELAI	0	0.138	0.967	0.242	0.140	0.073	5	0.984	0.238
RELAI	1	0.114	0.993	0.205	0.146	0.086	5	0.984	0.202
RELAI	2	0.488	0.276	0.353	0.087	0.082	2	0.996	0.352
RELAI	3	0.207	0.875	0.335	0.079	0.056	2	0.996	0.334
RELAI	4	0.119	0.868	0.209	0.120	0.089	2	0.996	0.208
CeDRI	0	0.110	0.993	0.199	0.109	0.090	2	0.996	0.198
CeDRI	1	0.116	**1**	0.207	0.113	0.085	2	0.996	0.206
CeDRI	2	0.105	**1**	0.190	0.096	0.094	**1**	**1**	0.190
BLUE	0	0.283	0.934	0.435	0.084	0.041	5	0.984	0.428
BLUE	1	0.142	0.875	0.245	0.117	0.081	4	0.988	0.242
BLUE	2	0.454	0.849	0.592	0.079	0.037	7	0.977	0.578
BLUE	3	0.394	0.868	0.542	0.075	0.035	5	0.984	0.534
BLUE	4	0.249	0.928	0.393	0.085	0.044	4	0.988	0.388
UPV-Symanto	0	0.307	0.678	0.422	0.097	0.051	5	0.984	0.416
UPV-Symanto	1	0.276	0.638	0.385	**0.059**	0.056	**1**	**1**	0.385
UPV-Symanto	2	0.313	0.645	0.422	0.072	0.053	2	0.996	0.420
UPV-Symanto	3	0.301	0.770	0.433	0.089	0.044	5	0.984	0.426
UPV-Symanto	4	0.198	0.711	0.310	0.082	0.063	3	0.992	0.307
UNSL	0	0.336	0.914	0.491	0.125	**0.034**	11	0.961	0.472
UNSL	1	0.110	0.987	0.198	0.093	0.092	**1**	**1**	0.198
UNSL	2	0.129	0.934	0.226	0.098	0.085	**1**	**1**	0.226
UNSL	3	0.464	0.803	0.588	0.064	0.038	3	0.992	0.583
UNSL	4	0.532	0.763	**0.627**	0.064	0.038	3	0.992	**0.622**

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
NLP-UNED	0	0.8	0.82	0.47	0.8	0.74	0.47	0	0	0	0	0	0
NLP-UNED	1	0.7	0.68	0.39	0.8	0.86	0.55	0	0	0	0	0	0
NLP-UNED	2	0.9	0.81	0.39	0.6	0.44	0.44	0	0	0	0	0	0
NLP-UNED	3	0.6	0.6	0.37	0.6	0.58	0.47	0	0	0	0	0	0
NLP-UNED	4	0.5	0.47	0.32	**0.9**	**0.94**	0.55	0	0	0	0	0	0
AvocadoToast	0	0	0	0.11	0.7	0.5	0.52	0	0	0	0	0	0
AvocadoToast	1	0	0	0.1	0.3	0.28	0.26	0	0	0	0	0	0
AvocadoToast	2	0.1	0.06	0.12	0.7	0.5	0.52	0	0	0	0	0	0
Birmingham	0	0.3	0.41	0.12	0	0	0	0	0	0	0	0	0
Birmingham	1	0	0	0.03	0	0	0	0	0	0	0	0	0
Birmingham	2	0.1	0.19	0.07	0	0	0	0	0	0	0	0	0
Birmingham	3	0.1	0.07	0.08	0	0	0	0	0	0	0	0	0
Birmingham	4	0	0	0.11	0	0	0	0	0	0	0	0	0
NuFAST	0	0	0	0.08	0	0	0	0	0	0	0	0	0
NuFAST	1	0	0	0.08	0	0	0	0	0	0	0	0	0
NuFAST	2	0	0	0.08	0	0	0	0	0	0	0	0	0
NaCTeM	0	0.1	0.06	0.07	0	0	0.08	0.1	0.06	0.15	0	0	0.06
NaCTeM	1	0.1	0.06	0.14	0.1	0.07	0.09	0	0	0.1	0.1	0.06	0.07
NaCTeM	2	0.2	0.19	0.15	0.2	0.19	0.11	0.1	0.06	0.13	0.1	0.06	0.09
NaCTeM	3	0.1	0.06	0.07	0.1	0.06	0.08	0	0	0.11	0.1	0.19	0.18
NaCTeM	4	0	0	0.07	0.1	0.06	0.06	0.1	0.06	0.1	0.1	0.06	0.08
EFE	0	0.5	0.35	0.37	0.8	0.74	0.63	0.8	0.74	0.6	0.8	0.81	0.62
EFE	1	0.5	0.35	0.37	0.8	0.74	0.63	0.8	0.74	0.6	0.8	0.81	0.62
EFE	2	0.7	0.68	0.49	0.5	0.44	0.56	0.6	0.55	0.59	0.6	0.55	0.59
EFE	3	0.7	0.68	0.49	0.5	0.44	0.56	0.6	0.55	0.59	0.6	0.55	0.59
BioInfo@UAVR	0	0.1	0.06	0.13	0	0	0	0	0	0	0	0	0
BioInfo@UAVR	1	0.1	0.06	0.07	0	0	0	0	0	0	0	0	0
NUS-IDS	0	0.8	0.86	0.55	0	0	0	0	0	0	0	0	0
NUS-IDS	1	0.8	0.75	0.49	0	0	0	0	0	0	0	0	0
NUS-IDS	2	0.9	0.81	0.49	0	0	0	0	0	0	0	0	0
NUS-IDS	3	0.6	0.73	0.46	0	0	0	0	0	0	0	0	0
NUS-IDS	4	0.8	0.85	0.52	0	0	0	0	0	0	0	0	0
RELAI	0	0.1	0.06	0.11	0.4	0.37	0.46	0.4	0.32	0.38	0.5	0.47	0.41
RELAI	1	0	0	0.12	0.2	0.12	0.36	0	0	0.27	0.1	0.06	0.28
RELAI	2	0.8	0.71	0.4	0.4	0.28	0.40	**1**	**1**	0.6	**1**	**1**	0.57
RELAI	3	0.7	0.76	0.43	0	0	0.31	0.9	0.88	0.59	0.8	0.75	0.56
RELAI	4	0.4	0.44	0.34	0	0	0.21	0.4	0.34	0.27	0.5	0.5	0.31
CeDRI	0	0.3	0.35	0.35	0.5	0.54	0.31	0	0	0	0	0	0
CeDRI	1	0.3	0.38	0.19	0.4	0.54	0.2	0	0	0	0	0	0
CeDRI	2	0.1	0.1	0.07	0.2	0.25	0.12	0	0	0	0	0	0
BLUE	0	0.7	0.75	0.54	0.8	0.82	0.59	0	0	0	0	0	0
BLUE	1	0.2	0.13	0.26	0.4	0.41	0.29	0	0	0	0	0	0
BLUE	2	0.6	0.49	0.50	**0.9**	**0.94**	0.55	0	0	0	0	0	0
BLUE	3	0.6	0.43	0.49	0.8	0.87	0.54	0	0	0	0	0	0
BLUE	4	0.7	0.61	0.52	0.8	0.88	0.55	0	0	0	0	0	0
UPV-Symanto	0	0.8	0.83	0.53	**0.9**	**0.94**	0.67	0.9	0.94	0.67	0	0	0
UPV-Symanto	1	0.8	0.88	0.5	0.8	0.69	0.64	0.8	0.69	0.64	0	0	0
UPV-Symanto	2	0.8	0.82	0.55	0.8	0.83	0.59	0.8	0.83	0.59	0	0	0
UPV-Symanto	3	0.6	0.70	0.51	**0.9**	**0.94**	0.69	0.9	0.94	0.69	0	0	0
UPV-Symanto	4	0.9	0.93	0.53	**0.9**	0.81	0.65	0.9	0.81	0.65	0	0	0
UNSL	0	**1**	**1**	**0.70**	0.7	0.74	**0.82**	0.8	0.81	0.8	0.8	0.81	**0.80**
UNSL	1	0.8	0.82	0.61	0.8	0.73	0.59	0.9	0.94	0.58	**1**	**1**	0.61
UNSL	2	0.3	0.27	0.28	0	0	0	0	0	0	0	0	0
UNSL	3	**1**	**1**	0.63	**0.9**	0.81	0.76	0.9	0.81	**0.71**	0.8	0.73	0.69
UNSL	4	**1**	**1**	0.63	**0.9**	0.81	0.76	0.9	0.81	**0.71**	0.8	0.73	0.69

4. Task 3: Measuring the Severity of the Signs of Depression

This task is a continuation of Task 3 from 2019 and Task 2 from 2020. The task consists of estimating the degree of depression based on a thread of user submissions. Participants were given the full history of postings for each user (in a single release of data), and they were required to fill out a standard depression questionnaire based on the evidence found in the history of postings. Participants in 2021 had the option of using 2019 and 2020 data as training data (filled questionnaires and social media submissions from the users, i.e. a training set composed of 90 users).
The questionnaire is derived from the Beck's Depression Inventory (BDI) [18], which assesses the presence of feelings like sadness, pessimism, loss of energy, etc, for the detection of depression. The questionnaire contains the 21 questions reported in Table 9.

Table 9: Beck's Depression Inventory

```
Instructions:

This questionnaire consists of 21 groups of statements.  Please read each group of statements
carefully, and then pick out the one statement in eachgroup that best describes the way you feel.
If several statements in thegroup 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.
```

Table 9: Beck's Depression Inventory (continued)

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.

16. Changes in Sleeping Pattern
 0. I have not experienced any change in my sleeping pattern.
 1a. I sleep somewhat more than usual.
 1b. I sleep somewhat less than usual.
 2a. I sleep a lot more than usual.

Table 9: Beck's Depression Inventory (continued)

```
 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.
 1a. My appetite is somewhat less than usual.
 1b. 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.
```

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 80 test users and they were asked to produce a file with the following structure:

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

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-).

4.1. Task 3: Evaluation Metrics

For consistency purposes, we employed the same evaluation metrics utilised in 2019 and 2020. 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 the same answer as the actual answers to the questionnaire. For example, an automatic questionnaire with five 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[7]
- **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)
  ```

[7] 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).

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.

4.2. Task 3: Results

Table 10 presents the results achieved by the participants in this task.
Starting with the AHR scores, the results in the task show that the best teams get rates below 40% of correct answers. These results do not improve but are aligned with the results obtained in the tasks of previous years (eRisk's Task 3 in 2019 and Task 2 in 2020), whose best AHR ratios were around 40%. This suggests that analyzing user posts can help extract some signals or symptoms related to depression. In the case of ACR, the best performing run (UPV-Symanto 4_symanto_upv_lingfeat_cor) shows a 73.17%, exceeding the 70% ACR barrier established in previous years, which represents a sustained improvement in the results of this metric for this task. However, this value is only slightly better than the näive all 1s algorithm (72.90%). This metric penalizes high distances between the correct answer and the answer given by the system and, thus, it somehow favours conservative answers. By always choosing 1, the all 1s algorithm sets an upper limit of the distance equal to 2 (it gets 2 when the correct answer is 3). In terms of AHR, some participating runs outperform the näive baseline algorithms (all 1s = 23.03%, all 0s = 32.02%). This implies that the distance-based ACR metric penalizes system failures in estimating response to an item more effectively.
These results put forth an existing barrier in the generalization process: from the specific estimation of individual answers (to each question in the questionnaire) to the overall estimation of the subject's depression level. In terms of ADODL, the best run (CYUT run 2) shows rates around 83.59%, representing a tiny percentage improvement compared to previous years (the best ADODL result obtained in Task 2 2020 was 83.15%).
Several teams offer values greater than 80% in the ADODL metric, strengthening the values obtained in previous years. However, the difficulty in the generalization process is clearly appreciated when we analyze the DCHR metric. In this case, the best performing run (CYUT run 2) gets the depression category right for only 41.25% of the individuals. This result is slightly lower than the maximum obtained in previous years (around 45% of individuals in Task 2 2020). This value is better than the baseline variants but, still, there is much room for improvement, and the trend in the data remains consistent throughout successive editions.
These results confirm the task's viability for automatically extracting some depression-related evidence from social media activity. Still, there is a need to improve the generalization process in order to advance towards a more comprehensive, more effective depression screening tool. Some of our future plans include to further analyze the participants' estimations (e.g., to determine which particular BDI questions are easier or harder to answer automatically) and to study whether or not specific questions of the questionnaire are more influential to the global depression score (ADODL and DCHR).

Table 10
Task 3 Results. Participating teams and runs with corresponding scores in AHR, ACR, ADODL and DCHR metrics. Stared runs did not submit decisions for every subject.

Run	AHR	ACR	ADODL	DCHR
BLUE run0	27.86%	64.66%	74.15%	17.50%
BLUE run1	30.00%	64.58%	70.65%	11.25%
BLUE run2	30.36%	65.42%	75.42%	21.25%
BLUE run3	29.52%	64.70%	73.63%	13.75%
BLUE run4	29.76%	65.04%	74.84%	15.00%
CYUT run1	32.02%	66.33%	75.34%	20.00%
CYUT run2	32.62%	69.46%	**83.59%**	**41.25%**
CYUT run3	28.39%	63.51%	80.10%	38.75%
DUTH_ATHENA MaxFT	31.43%	64.86%	74.46%	15.00%
DUTH_ATHENA MeanFT	32.02%	65.63%	73.81%	12.50%
DUTH_ATHENA MeanPosts	25.06%	63.97%	80.28%	30.00%
DUTH_ATHENA MeanPostsAB	33.04%	67.86%	80.32%	27.50%
DUTH_ATHENA MeanPostsSVM	**35.36%**	67.18%	73.97%	15.00%
NaCTeM run1	31.43%	64.54%	74.98%	18.75%
NaCTeM run2	31.55%	65.00%	75.04%	21.25%
NaCTeM run3	32.86%	66.67%	76.23%	22.50%
RELAI dmknn_dan	34.64%	67.58%	78.69%	23.75%
RELAI dmknn_danb	30.18%	65.26%	78.91%	25.00%
RELAI etm *	38.78%	72.56%	80.27%	35.71%
RELAI k_nn_dan	34.82%	66.07%	72.38%	11.25%
RELAI lda	28.33%	63.19%	68.00%	10.00%
Tanvi_Darci run 0	35.12%	67.76%	75.81%	22.50%
Unior_NLP uniorA	31.67%	63.95%	69.42%	08.75%
Unior_NLP uniorB	31.61%	64.66%	74.74%	15.00%
Unior_NLP uniorC	28.63%	63.31%	76.45%	20.00%
Unior_NLP uniorD	28.10%	64.25%	71.27%	15.00%
uOttawa1_sim_BERT_base+	28.39%	65.73%	78.91%	25.00%
uOttawa2_Top2Vec_USE+	28.04%	63.00%	77.32%	27.50%
uOttawa3_sim_BERT_large+	25.83%	59.68%	71.23%	27.50%
uOttawa4_Ensemble_BERT_QA	27.68%	62.08%	76.92%	20.00%
uOttawa5_sim_ROBERTA+	26.31%	62.60%	76.45%	30.00%
UPV-Symanto 0_symanto_upv_svm_linear_drb	34.58%	67.32%	75.62%	26.25%
UPV-Symanto 1_symanto_upv_svm_linear_mt30	32.20%	66.05%	77.28%	26.25%
UPV-Symanto 2_symanto_upv_svm_linear	33.15%	66.05%	75.42%	23.75%
UPV-Symanto 3_symanto_upv_rfc_df40_mt30	33.09%	66.39%	76.87%	23.75%
UPV-Symanto 4_symanto_upv_lingfeat_cors	34.17%	**73.17%**	82.42%	32.50%
All 0s Baseline	23.03%	54.92%	54.92%	7.50%
All 1s Baseline	32.02%	72.90%	81.63%	33.75%

5. Participating Teams

Table 11 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.

Table 11
eRisk 2021 participants

team	Task 1 #runs	Task 2 #runs	Task 3
RELAI	5	5	5
UPV-Symanto	5	5	5
BLUE	5	5	5
UNSL	5	5	
CEDRI	2	3	
EFE	4	4	
NLP-UNED		5	
AvocadoToast		3	
Birmingham		5	
NuFAST		3	
BioInfo@UAVR		2	
NUS-IDS		5	
NAcTem		5	3
CYUT			3
DUTH-ATHENA			3
Tanvi-Darci			1
Unior-NLP			4
uOttawa			5

RELAI [19]. The team of the Université du Québec à Montréal (Canada). Regarding T1, the team approach includes some external and existing testimonials' data and self-evaluation questionnaire results for a distance-based model. For T2, two approaches based on neural networks were tested. One is based on the Contextualizer encoder, while the other is based on RoBERTa embeddings. Finally, the team presented a similarity-based model for T3, computing the similarities of the writings of the test subjects to those of the training subjects. These representations were based on topic modelling or neural encoders trained on authorship decisions. The team also tested some variations on similarity-based regression and a nearest-neighbours approach.
UPV-Symanto [20]. This team is composed of researchers from Universitat Politécnica de Valencia, Symanto Research, and from the University of Bucharest. The team participated in the three tasks. Task 1 and Task 2 collected external information from Reddit (800 users T1 and 46 users from T2). For T1, they trained a Transformer-based classifier over Bert. Task 2 repeated the same approach as in Task 1 but also tested Hierarchical Attention Networks with different linguistic features (GloVe embeddings, style, LIWC, etc.). Finally, for T3, they tested two approaches. The first is based on a temporal user representation based on the evolution of some of the linguistics features over time. In the second one, the UPV-Symanto team trains a classifier based on RoBRoBERTaERTA over external data from Reddit. Then they use the CLS embeddings from the user posts (averaged) for training 21 classifiers (one per question)
BLUE [21]. This is a team from the University of Bucharest (Romania). They present models using BERT transformers and automatic data crawling from mental health subreddits on all three tasks. They follow a data acquisition approach similar to last years' iLab team. They used

the same model for producing the runs but different training data (different combinations of eRisk Training data and crawled data) or thresholds for the different tasks.

UNSL [22]. This team is a collaboration between Universidad Nacional de San Luis (UNSL) and the Instituto de Matemática Aplicada San Luis (IMASL), both in Argentina. The team proposed a general overview of both T1 and T2 focusing on the delay aspect of the early risk detection problem called Early Risk Detection Framework (ERD), that explicitly identifies how the classification with partial information (CPI) component, which is the risky-user classification model is implemented, and how the deciding the moment of classification (DMC) component makes its decisions (the early alert policy). Applying the general framework, the team presented different feature variants and configurations of the parameters of the ERD framework to address T1. For t2, the team presented five different models, including `doc2vec` variants and configurations of the parameters of the ERD framework.

CEDRI [23]. The team from the Research Center for Digitalization and Intelligent Robotics (CeDRI), at Instituto Politécnico de Bragança, (Portugal) participated in T1 and T2. For T1, the team contribution is focused on creating new training data in a post-level automatic annotation effort. With that data, the team used logistic regressors, CNNs, and LSTM classifiers. Regarding the self-harm task, this group employed the provided training data. Still, selected submissions to focus on the contents that are more related to self-harm (based on a non-suicidal self-injury vocabulary) and employed similar classifiers.

NLP_UNED team [24]. This team is a collaboration between the Universidad Nacional de Educación a Distancia (UNED) and the Instituto Mixto de Investigación of Escuela Nacional de Sanidad (IMIENS), Spain. They present results for T2. The proposed model uses a combination of text-based (lexical self-harms vocabularies, grammatical aspects, and sentiment analysis) and TF-IDF based (term frequency-inverse document frequency) features in an SVM classifier to predict whether a message belongs to a positive or negative user in self-harm. They present three stages: data pre-processing, feature calculation, message classification.

Birmingham team [24]. The team from the Center for Computational Biology, University of Birmingham, UK) participated in T2. This group focused on standard (and mostly classic) machine learning methods (testing AdaBoost, Logistic Regression, Random Forest and SVM), together with standard weighting schemes and feature selection techniques such as bag of words or `doc2vec` based features.

NUFAST [25]. This is a team from the National University of Computer Emerging Science National Center for Text Mining, Karachi, Pakistan. They implemented different classifiers based on Logistic Regression and using BERT embeddings as document representation.

BioInfo@UAVR [26]. The team from the ETI/IEETA, University of Aveiro, Portugal. The group experimented in two directions for T2. Firstly, with some BERT-based solutions with an SVM classifier. Secondly, adding to the previous model some variants based on sentiment analysis. They used VADER, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

NaCTeM [27]. This is a team from the National Center for Text Mining, University of Manchester, United Kingdom. They implemented a host of models for Task 3 based on neural networks. They experimented with pre-trained language models for feature extraction from all user's posts (ELMo, BERT, and SpanEmo). Then, they trained a random forest classifier on top of that extracted features. They also evaluated different layers in the SpanEmo method determining

which one is best for obtaining better results.
CYUT [28]. This is a team from the Chaoyang University of Technology, Taiwan. They approached Task 3. For that, they employed a RoBERTa pre-trained model for each BDI item to classify all the posts of each user. They tested three different approaches to obtain a user-level estimation from post-level from the post-level predictions obtaining the final estimations.
DUTH-ATHENA [29]. From Democritus University of Thrace and Athena Research Center, this team, Xanthi, Greece, implemented three different approaches to fill the BDI questionnaire (Task 3) automatically. In the first two approaches, they use a sentence-based language model, called SBERT, to represent the subjects' posts and employ cosine similarity and well-known classifiers, such as SVM and Random Forests, to classify each user in the options of each BDI item. In the third approach, they fine-tune a RoBERTa model to estimate the answers of the subjects of the collection.
uOttawa [30]. This the team from the School of Electrical Engineering and Computer Science, University of Ottawa, Canada. They presented several deep learning methods for Task 3. They perform a pre-filtering process based on the relevance of posts on the data. After that, they experimented with various classification variants, including transfer learning techniques (Zero-shot learning) and Question Answering (QA) systems, such as BERT and Universal Sentence Encoder QA.
Unior-NLP [31]. This is a team from "L' Orientale", University of Naples, Italy. They present the results obtained by applying several steps of text pre-processing and feature extraction and two variants for feature representation (Latent Dirichlet Allocation and a general-purpose pre-trained model from Sentence Transformers library) to get input data for traditional machine learning classifiers.

6. Conclusions

This paper provided an overview of eRisk 2021. The fifth edition of this lab focused on two types of tasks. On the one hand, two tasks were on early detection of pathological gambling and self-harm (Task 1 and 2, respectively), where participants had sequential access to the user's social media posts and had to send alerts about at-risk individuals. On the other hand, one task was released to measuring the severity of the signs of depression (Task 3), 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
The proposed tasks received 117 runs from 18 teams in total. Although the effectiveness of the proposed solutions is still limited, the experimental results show that evidence extracted from social media is valuable, and automatic or semi-automatic screening tools could be developed to detect at-risk individuals. These results encourage us to further investigate the development of benchmarks for text-based screening of risk indicators.

Acknowledgements

This work was supported by projects 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, ED431B 2019/03) 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. URL: https://doi.org/10.1007/978-3-030-15719-7_35. doi:10.1007/978-3-030-15719-7_35.

[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. URL: https://doi.org/10.1007/978-3-030-58219-7_20. doi:10.1007/978-3-030-58219-7_20.

[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. URL: http://ceur-ws.org/Vol-2696/paper_253.pdf.

[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. URL: https://doi.org/10.1007/978-3-030-45442-5_72. doi:10.1007/978-3-030-45442-5_72.

[12] 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. URL: https://doi.org/10.1007/978-3-030-72240-1_76. doi:10.1007/978-3-030-72240-1_76.

[13] 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.

[14] 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. URL: http://ceur-ws.org/Vol-2621/CIRCLE20_23.pdf.

[15] 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. URL: https://doi.org/10.1145/3412841.3441947. doi:10.1145/3412841.3441947.

[16] 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).

[17] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: WSDM, ACM, 2018, pp. 495–503.

[18] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An Inventory for Measuring Depression, JAMA Psychiatry 4 (1961) 561–571.

[19] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas, M.-J. Meurs, Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.

[20] A. Basile, M. Chinea-Rios, A.-S. Uban, T. Müller, L. Rössler, S. Yenikent, B. Chulví, P. Rosso, M. Franco-Salvador, Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.

[21] A.-M. Bucur, A. Cosma, L. Dinu, Early risk detection of pathological gambling, self-harm

and depression using bert, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[22] J. M. Loyola, S. Burdisso, H. Thompson, L. Cagnina, M. Errecalde, Unsl at erisk 2021: A comparison of three early alert policies for early risk detection, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[23] R. P. Lopes, Cedri at erisk 2021: A naive approach to early detection of psychological disorders in social media, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[24] E. Campillo-Ageitos, H. Fabregat, L. Araujo, J. Martinez-Romo, Nlp-uned at erisk 2021: self-harm early risk detection with tf-idf and linguistic features, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[25] Q. U. Nisa, R. Muhammad, Towards transfer learning using bert for early detection of self-harm of social media users, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[26] A. Trifan, L. Ribeiro, J. L. Oliveira, Vader meets bert: sentiment analysis for early detection of signs of self-harm through social mining, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[27] H. Alhuzali, T. Zhang, S. Ananiadou, Predicting sign of depression via using frozen pre-trained models and random forest classifier, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[28] S.-H. Wu, Z.-J. Qiu, A roberta-based model on measuring the severity of the signs of depression, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[29] C. Spartalis, G. Drosatos, A. Arampatzis, Transfer learning for automated responses to the bdi questionnaire, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[30] D. Inkpen, R. Skaik, P. K. Gamaarachchige, D. Angelov, M. T. Fredenburgh, uottawa at erisk 2021: Automatic filling of the beck's depression inventory questionnaire using deep learning, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
[31] R. Manna, J. Monti, Unior nlp at erisk 2021: Assessing the severity of depression with part of speech and syntactic features, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.