Overview of eRisk at CLEF 2023: Early Risk Prediction on the Internet (Extended Overview)

Notebook for the eRisk Lab at CLEF 2023

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 article provides an overview of eRisk 2023, the seventh edition of the CLEF conference's lab dedicated to early risk detection. Our lab has been committed to exploring evaluation methodologies, effectiveness metrics, and other associated processes in the field of early risk detection since its inception. The applications of early alerting models are wide-ranging and span various domains, including health and safety. eRisk 2023 encompassed three tasks. The initial task involved ranking sentences based on their relevance to standardized depression symptoms. The second task concentrated on detecting signs associated with pathological gambling early. Lastly, the third task required participants to automatically estimate an eating disorders questionnaire by analyzing user writings on social media. In this extended overview, we include additional details about the participants' proposals and more detailed explanations about metrics.

Keywords

Early risk, Depression, Pathological gambling, Eating disorders

1. Introduction

The main objective of eRisk is to investigate evaluation methodologies, metrics, and other pertinent factors about research collection development and identifying signs associated with early risk detection. The potential of early detection technologies is significant, particularly in fields that address safety and health applications. Automated systems can play a crucial role in issuing early warnings in scenarios involving individuals displaying symptoms of mental

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

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)

CEUR Workshop Proceedings (CEUR-WS.org)

illnesses, infants encountering interactions with sexual abusers, or potential criminals publishing antisocial threats online.

Our lab primarily focuses on psychological issues, specifically depression, self-harm, pathological gambling, and eating disorders. Through our work, we have found that the relationship between psychological diseases and language use is intricate, and there is room for enhancing the effectiveness of automatic language-based screening models. In 2017, we conducted an exploratory task on early detection of depression, employing innovative evaluation methods and a test dataset described in [1, 2]. The following year we continued to foster the detection of early signs of depression and introduced a new task for detecting early signs of anorexia [3, 4]. In 2019, we further expanded the challenge by focusing on the early identification of anorexia symptoms, introducing a new challenge on early detection of self-harm, and proposing a third task centred around estimating a user's responses to a depression questionnaire based on their social media interactions [5, 6, 7]. In 2020, our efforts continued with the early detection of self-harm and introducing a task for estimating the severity of depression symptoms [8, 9, 10]. In 2021, our focus shifted to two scenarios: early detection of pathological gambling and self-harm and a task for severity estimation of depression [11, 12, 13]. Lastly, in the previous year, we presented three tasks: early pathological gambling detection, early detection of depression, and severity estimation of eating disorders [14, 15, 16].

In 2023, eRisk featured three campaign-style tasks [17, 18]. The first task involved ranking sentences based on their relevance to each of the 21 symptoms of depression derived from the BDI-II questionnaire. Participants in task 1 were provided with a collection of sentences extracted from publications of social media users. The second task represented the third edition of early risk detection for pathological gambling. The third task marked the second edition of the eating disorder severity estimation challenge. Detailed descriptions of these tasks can be found in the subsequent sections of this overview article.

A total of 98 teams registered for the lab, out of which we received results from 20 teams, with 37 runs for Task 1, 48 runs for Task 2, and 20 runs for Task 3.

2. Task 1: Search for Symptoms of Depression

Task 1 introduced a novel challenge in 2023 involving the production of sentence rankings based on their relevance to specific symptoms of depression. We instructed participants to rank sentences derived from user writings according to their relevance to the 21 standardized symptoms outlined in the BDI-II Questionnaire [19]. In this context, a sentence was deemed relevant to a particular symptom if it provided information about the user's state of that symptom. It is important to emphasize that a phrase could be considered relevant even if it conveyed positive information about the symptom. For instance, "I feel quite happy lately" should be regarded as relevant for symptom 1, "Sadness," in the BDI-II.

2.1. Dataset and format

The corpus provided to the participants was a TREC formatted sentence-tagged dataset (based on eRisk's past data). Table 1 reports some statistics of the corpus.

Table 1
Corpus statistics for Task 1: Search for Symptoms of Depression.

Number of users	3,107
Number of sentences	3,807,115
Average number of words per sentence	13.63

```
1 Q0 sentence-id-121 0001 10 myGroupNameMyMethodName
1 Q0 sentence-id-234 0002 9.5 myGroupNameMyMethodName
1 Q0 sentence-id-345 0003 9 myGroupNameMyMethodName
...
21 Q0 sentence-id-456 0998 1.25 myGroupNameMyMethodName
21 Q0 sentence-id-242 0999 1 myGroupNameMyMethodName
21 Q0 sentence-id-347 1000 0.9 myGroupNameMyMethodName
```

Figure 1: Example of TREC format for a participant's run for Task 1.

Table 2
Task 1 (Search for Symptoms of Depression): Number of runs from participants.

Team	# of submissions
BLUE	5
Formula-ML	4
GMU-FAST	2
Mason-NLP	1
NailP	5
OBSER-MENH	5
RELAI	5
UMU	2
UNSL	3
uOttawa	5
Total	37

Given the corpus of sentences and the description of the symptoms from the BDI questionnaire, the participants were free to decide on the best strategy to derive queries for representing the BDI symptoms. Each participating team submitted up to 5 variants (runs). Each run included 21 TREC-style formatted rankings of sentences, as shown in Figure 1.

2.2. Assessment Process

For each symptom, the participants could should submit up to 1000 results sorted by estimated relevance. We received 37 runs from 10 participating teams (see Table 2).
The generation of relevance judgments involved the participation of three expert assessors who annotated a pool of sentences associated with each symptom. We selected the candidate sentences using a top-k pooling process using the participants' submissions (37 different ranking methods)

The assessors were provided with explicit instructions regarding the determination of sentence relevance. They were instructed to consider a sentence relevant if it pertained to the symptom and provided explicit information about the individual's state in relation to it. This dual concept of relevance, encompassing both the topic and the reflection of the user's state, introduced a higher complexity level than more conventional relevance assessments. As a result, we developed a robust annotation methodology and formal assessment guidelines to ensure consistency and accuracy. To create the pool of sentences for assessment, we employed $k = 50$ in the pooling method. Table 3 presents the resulting pool sizes per sentence. It was observed that certain sentences had identical text but different IDs, potentially stemming from multiple users writing the same content. To reduce the assessors' workload, we automated removing duplicate sentences. The annotation software, specifically developed to support the annotation process of eRisk 2023, automatically assigned the assessors' relevance labels to all identical sentences. Despite implementing these optimizations, the average time spent per sentence by the three assessors was still over 30 seconds, even for this sentences that are very short documents. These resulted in more than 210 hours of assessors' time.

Table 3
Task 1 (Search for Symptoms of Depression): Size of the pool for every BDI Item

BDI Item (#)	original	unique	# rels (3/3)	# rels (2/3)
Sadness (1)	1110	1069	179	318
Pessimism (2)	1150	1096	104	325
Past Failure (3)	973	918	160	300
Loss of Pleasure (4)	1013	948	97	204
Guilty Feelings (5)	829	794	83	143
Punishment Feelings (6)	1079	1036	21	50
Self-Dislike (7)	1005	957	158	288
Self-Criticalness (8)	1072	1023	76	174
Suicidal Thoughts or Wishes (9)	953	923	260	349
Crying (10)	983	917	230	320
Agitation (11)	1080	1057	69	154
Loss of Interest (12)	1077	1021	70	168
Indecisiveness (13)	1110	1044	61	141
Worthlessness (14)	1067	986	71	144
Loss of Energy (15)	1082	1027	129	204
Changes in Sleeping Pattern (16)	938	904	203	350
Irritability (17)	1047	1008	94	155
Changes in Appetite (18)	984	947	103	224
Concentration Difficulty (19)	1024	981	83	141
Tiredness or Fatigue (20)	1033	994	123	221
Loss of Interest in Sex (21)	971	922	97	158

The annotation process involved a team of three assessors with diverse backgrounds and expertise. One assessor was a psychologist, while the other two were computer science researchers—a postdoctoral fellow and a PhD student—specialised in early-risk technologies.

The lab organisers conducted a preparatory session with the assessors to ensure consistency and clarity. During this session, an initial version of the guidelines was discussed, and any doubts or questions raised by the assessors were thoroughly addressed. Through this collaborative effort, the final version of the guidelines shown in Figure 2 was developed.

```
Assume you are given a BDI item, e.g.:
15. Loss of Energy
-I have as much energy as ever.
-I have less energy than I used to have.
-I don’t have enough energy to do very much.
-I don’t have enough energy to do anything.

The task consists of annotating sentences in the collection that are
    topically-relevant to the item (related to the question and/or to the
    answers).

Note: A relevant sentence should provide some information about the state of
    the own writer related to the topic of the BDI item. But it is not
    necessary that the exact same words are used.

Your job is to assess sentences on how topically-relevant they are for a
    concrete BDI item.
The relevance grades are:
1. Relevant: A relevant sentence should be topically-related to the BDI-item
    (regardless of the wording) and, additionally, it should refer to the
    state of the writer about the BDI-item.
0. Non-relevant: A non-relevant sentence does not address any topic related
    to the question and/or the answers of the BDI-item (or it is related to
    the topic but does not represent the writer’s state about the BDI-item).
    For example, for BDI-item 15, a sentence that does not talk about the
    individual’s level of energy (regardless of the wording), then is a non-
    relevant sentence.

Examples (assessment of sentences ranked for BDI-item number 15):
I cannot control my energy these days: Relevant
My sister has no energy at all: Non-relevant sentence (because it does not
    refer to the writer who posted this sentence)
The book was about a highly energetic man: Non-relevant sentence (because it
    does not refer to the writer who posted this sentence)
I feel more tired than usual: Relevant
The football team is named Top Energy: Non-relevant
I am totally lonely: Non-relevant (it does not mention energy)
I have just recharged my batteries: Relevant
I am lost: Non-relevant

We advise you to not stop the assessment session in the middle of one BDI-
    item (this helps to maintain consistency in the judgments). To measure
    the assessment effort, we ask you to record the time spent on fully
    evaluating the sentences presented for each BDI-item.
```

Figure 2: Guidelines for labelling sentences related to depression symptoms (Task 1).

Following these guidelines, a sentence is deemed relevant only if it provides "some information about the state of the individual related to the topic of the BDI item." This criterion serves as the foundation for determining the relevance of sentences during the annotation process.

After the initial meeting, the assessors proceeded to label the pools of sentences for the first three BDI topics (sadness, pessimism, and past failure). Following this phase, we organized another meeting to address any additional concerns or questions that arose during the annotation process. This collaborative session was vital in refining the annotation criteria and ensuring consistency. The final outcomes of the annotation process, including the number of relevant sentences per BDI item, are presented in Table 3 (last two columns). We applied two aggregation criteria for sentence relevance determination: unanimity and majority.

The performance results of the participating systems are presented in Tables 4 (majority-based qrels) and 5 (unanimity-based qrels). These tables show results for multiple standard ranking metrics, namely Mean Average Precision (MAP), mean R-Precision, mean Precision at 10, and mean NDCG at 1000. Notably, Formula-ML models based on setence transformers, developed by the NITK Surathkal team, achieved the highest performance rankings across all metrics and relevance judgment types.

3. Task 2: Early Detection of Pathological Gambling

This task marks the third edition of the challenge, which aims to develop innovative models for the early identification of pathological gambling risk. Pathological gambling, also known as ludomania or "gambling addiction," involves an irresistible urge to gamble despite the adverse consequences. According to the World Health Organization (WHO), the prevalence of adult gambling addiction in 2017 ranged from 0.1% to 6.0% [29]. The objective of this task was to process evidence in a sequential manner and detect early indications of compulsive or disordered gambling as promptly as possible. Participating systems were required to analyze user posts on social media in the order of their occurrence. Successful outcomes from this task could potentially be employed for sequential monitoring of user interactions across diverse online platforms such as blogs, social networks, and other forms of digital media.

The test collection employed for this task followed the same format as the collection described in the work by Losada and Crestani [30]. This collection comprises writings, including posts and comments, obtained from a carefully selected group of social media users. Within this dataset, users are classified into two categories: pathological gamblers and non-pathological gamblers. For each user, the collection contains a chronological sequence of writings. To facilitate the task and ensure equitable distribution, we established a dedicated server that systematically provided user writings to the participating teams. Further details regarding the setup and functioning of the server can be found on the lab's official website[1].

This task followed a train-test approach. During the training stage, teams were provided with training data, including the complete history of writings for training users. We indicated which users explicitly identified themselves as pathological gamblers, allowing participants to tune their systems using this training data. In 2023, the training data for Task 1 consisted of users from previous editions of the self-harm task.

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

Table 4
Ranking-based evaluation for Task 1 (majority voting)

Team	Run	AP	R-PREC	P@10	NDCG
Formula-ML [20]	SentenceTrainsformers_0.25	**0.319**	**0.375**	**0.861**	**0.596**
Formula-ML	SentenceTrainsformers_0.1	0.308	0.359	**0.861**	0.584
Formula-ML	result2	0.086	0.170	0.457	0.277
Formula-ML	word2vec_0.1	0.092	0.176	0.500	0.285
OBSER-MENH [21]	salida-distilroberta-90-cos	0.294	0.359	0.814	0.578
OBSER-MENH	salida-mpnet-90-cos	0.265	0.333	0.805	0.550
OBSER-MENH	salida-mpnet-21-cos	0.120	0.207	0.471	0.365
OBSER-MENH	salida-distilroberta-21-cos	0.158	0.249	0.543	0.418
OBSER-MENH	salida-mini12-21-cos	0.114	0.184	0.305	0.329
uOttawa [22]	USESim	0.160	0.248	0.600	0.382
uOttawa	Glove100Sim	0.017	0.052	0.195	0.105
uOttawa	RobertaSim	0.033	0.080	0.329	0.150
uOttawa	GloveSim	0.011	0.038	0.162	0.075
uOttawa	BertSim	0.084	0.150	0.505	0.271
BLUE [23]	SemSearchOnBDI2Queries	0.104	0.126	0.781	0.211
BLUE	SemSearchOnGeneratedQueriesMentalRoberta	0.029	0.063	0.367	0.105
BLUE	SemSearchOnBDI2QueriesMentalRoberta	0.027	0.044	0.386	0.089
BLUE	SemSearchOnGeneratedQueries	0.052	0.074	0.586	0.139
BLUE	SemSearchOnAllQueries	0.065	0.086	0.629	0.160
NailP [24]	T1_M2	0.095	0.146	0.519	0.226
NailP	T1_M4	0.095	0.146	0.519	0.221
NailP	T1_M3	0.073	0.114	0.471	0.180
NailP	T1_M5	0.089	0.140	0.486	0.223
NailP	T1_M1	0.074	0.114	0.471	0.189
RELAI [25]	bm25|mpnetbase	0.048	0.081	0.538	0.140
RELAI	BM25	0.016	0.061	0.043	0.145
RELAI	bm25|mpnetbase_simcse	0.030	0.066	0.390	0.114
RELAI	bm25|mpnetqa_simcse	0.027	0.063	0.376	0.109
RELAI	bm25|mpnetqa	0.038	0.075	0.438	0.126
UNSL [26]	Prompting-Classifier	0.036	0.090	0.229	0.180
UNSL	Similarity-AVG	0.001	0.008	0.010	0.016
UNSL	Similarity-MAX	0.001	0.011	0.019	0.019
UMU [27]	LexiconMultilingualSentenceTransformer	0.073	0.140	0.495	0.222
UMU	LexiconSentenceTransformer	0.054	0.122	0.362	0.191
GMU-FAST	FAST-DCMN-COS-INJECT	0.001	0.002	0.014	0.004
GMU-FAST	FAST-DCMN-COS-INJECT_FULL	0.001	0.003	0.014	0.005
Mason-NLP [28]	MentalBert	0.035	0.072	0.286	0.117

During the test stage, participants connected to our server and engaged in an iterative process of receiving user writings and sending their responses. Participants had the discretion to pause the process and issue an alert at any point in the chronology of user writings. After reading each user's writing, teams had to decide between two options: i) issuing an alert about the user, indicating a predicted sign of gambling risk, or ii) not issuing an alert. Each participant independently made this choice for every user in the test split. It is important to note that once an alert was issued, it was considered final, and no further decisions regarding that particular individual were taken into account. On the other hand, we did not consider the absence of

Table 5
Ranking-based evaluation for Task 1 (unanimity)

Team	Run	MAP	R-PREC	P@10	NDCG
Formula-ML [20]	SentenceTransformers_0.25	0.268	**0.360**	**0.709**	**0.615**
Formula-ML	SentenceTransformers_0.1	**0.293**	0.350	0.685	0.611
Formula-ML	result2	0.079	0.155	0.357	0.290
Formula-ML	word2vec_0.1	0.085	0.163	0.357	0.299
OBSER-MENH [21]	salida-distilroberta-90-cos	0.281	0.344	0.652	0.604
OBSER-MENH	salida-mpnet-90-cos	0.252	0.337	0.643	0.575
OBSER-MENH	salida-distilroberta-21-cos	0.135	0.216	0.390	0.413
OBSER-MENH	salida-mini12-21-cos	0.099	0.165	0.214	0.329
OBSER-MENH	salida-mpnet-21-cos	0.101	0.189	0.319	0.366
uOttawa [22]	USESim	0.139	0.232	0.438	0.380
uOttawa	GloveSim	0.008	0.028	0.110	0.063
uOttawa	Glove100Sim	0.011	0.042	0.110	0.092
uOttawa	RobertaSim	0.025	0.068	0.190	0.140
uOttawa	BertSim	0.070	0.130	0.357	0.260
BLUE [23]	SemSearchOnBDI2Queries	0.129	0.167	0.643	0.260
BLUE	SemSearchOnAllQueries	0.067	0.105	0.452	0.177
BLUE	SemSearchOnGeneratedQueriesMentalRoberta	0.018	0.059	0.186	0.085
BLUE	SemSearchOnGeneratedQueries	0.052	0.088	0.381	0.147
BLUE	SemSearchOnBDI2QueriesMentalRoberta	0.032	0.058	0.300	0.104
NailP [24]	T1_M2	0.090	0.143	0.410	0.229
NailP	T1_M4	0.090	0.143	0.410	0.224
NailP	T1_M5	0.083	0.139	0.338	0.222
NailP	T1_M1	0.073	0.114	0.343	0.192
NailP	T1_M3	0.073	0.114	0.343	0.181
UMU [27]	LexiconSentenceTransformer	0.044	0.110	0.210	0.175
UMU	LexiconMultilingualSentenceTransformer	0.059	0.125	0.333	0.209
RELAI [25]	BM25	0.012	0.036	0.019	0.135
RELAI	bm25|mpnetbase_simcse	0.026	0.059	0.243	0.103
RELAI	bm25|mpnetqa_simcse	0.023	0.052	0.262	0.097
RELAI	bm25|mpnetqa	0.030	0.065	0.290	0.109
RELAI	bm25|mpnetbase	0.039	0.069	0.343	0.124
UNSL [26]	Similarity-MAX	0.001	0.006	0.010	0.012
UNSL	Prompting-Classifier	0.020	0.063	0.090	0.157
UNSL	Similarity-AVG	0.000	0.005	0.005	0.011
GMU-FAST	FAST-DCMN-COS-INJECT_FULL	0.001	0.003	0.014	0.006
GMU-FAST	FAST-DCMN-COS-INJECT	0.001	0.002	0.010	0.003
Mason-NLP [28]	MentalBert	0.024	0.054	0.190	0.099

alerts as final, allowing participants to subsequently submit an alert if they detected emerging signs of risk.

In constructing the ground truth assessments, we employed established approaches that aim to optimize the utilization of assessors' time [31, 32]. These methods utilize simulated pooling strategies, which facilitate the efficient creation of test collections. The key statistics of the test collection used for Task 2 are presented in Table 6. .

To evaluate the performance of the systems, we used two indicators: the accuracy of the decisions made and the number of user writings required to reach those decisions. These criteria offer

Table 6
Task 2 (pathological gambling). Main statistics of test collection

	Pathological Gamblers	*Control*
Num. subjects	103	2071
Num. submissions (posts & comments)	33,719	1,069,152
Avg num. of submissions per subject	327.33	516.25
Avg num. of days from first to last submission	$\approx$ 675	$\approx$ 878
Avg num. words per submission	28.9	20.47

valuable insights into the evaluated systems' effectiveness and efficiency. To support the test stage, the server distributed user writings iteratively and waited for responses from participants. Importantly, new user data was only provided to a specific participant after the service received the response from previous evidence from that particular team. The submission period for the task was open from January 16th, 2023, until April 14th, 2023.

3.1. 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 [30]. 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 [33] 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

[2] computed with respect to the positive class.

$ERDE_o^{\%}$. 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 [34]. 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):

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

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

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

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

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

where p is a parameter that determines how quickly the penalty should increase. In [34], p was set such that the penalty equals 0.5 at the median number of posts of a user[5]. Observe that a

[3] Observe that Sadeque et al (see [34], 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.

decision right after the first writing has no penalty (i.e. $penalty(1) = 0$). Figure 3 plots how the latency penalty increases with the number of observed writings.

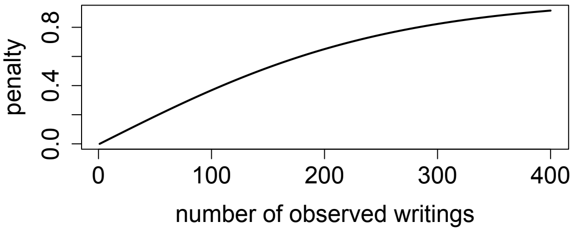

Figure 3: Latency penalty increases with the number of observed writings (k_u)

The system's overall speed factor is computed as:

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

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:

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

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 $ERDE$.

3.2. Ranking-based Evaluation

In addition to the evaluation discussed above, we employed an alternative form of evaluation to further assess the systems. After each data release (new user writing), participants were required to provide the following information for each user in the collection:

- A decision for the user (alert or no alert), which was used to calculate the decision-based metrics discussed previously.

Table 7
Task 2 (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)
UNSL	3	2004	1 day 02:17:36.417
ELiRF-UPV	1	2004	1 day 13:03:54.419
Xabi_EHU	5	2004	4 days 23:52:41.454
OBSER-MENH	5	2004	6 days 03:56:44.247
RELAI	5	764	6 days 09:12:00.148
NLP-UNED-2	5	2004	7 days 04:24:33.158
NUS-eRisk	5	2004	9 days 14:39:26.347
BioNLP-IISERB	5	61	10 days 00:49:40.529
SINAI	5	809	10 days 13:00:02.164
UMU	5	2004	14 days 00:29:30.434
NLP-UNED	5	1151	54 days 19:27:42.538

- A score representing the user's level of risk, estimated based on the evidence observed thus far.

The scores were used to create a ranking of users in descending order of estimated risk. For each participating system, a ranking was generated at each data release point, simulating a continuous re-ranking approach based on the observed evidence. In a real-life scenario, this ranking would be presented to an expert user who could make decisions based on the rankings (e.g., by inspecting the top of the rankings). Each ranking can be evaluated using standard ranking metrics such as P@10 or NDCG. Therefore, we report the performance of the systems based on the rankings after observing different numbers of writings

3.3. Results

Table 7 presents the participating teams, the number of runs submitted, and the approximate time duration from the first to the last response. This time-lapse indicates the level of automation and efficiency achieved by each team's algorithms. While a few submitted runs processed the entire thread of messages (2004), many variants terminated earlier. It is noteworthy that some teams were still submitting results at the deadline. Two teams demonstrated relatively fast processing times, taking approximately a day to analyze the complete history of user messages. In contrast, the remaining teams required several days to complete the entire process.
Table 8 reports the decision-based performance achieved by the participating teams. In terms of Precision, $F1$, $ERDE_5$, $ERDE_{50}$, and latency-weighted $F1$ the best performing team was the ELiRF-UPV (run 0). Regarding $latency_{TP}$ and $speed$ SINAI (runs 0 and 2) are the ones that having perfect values obtained the best $F1$. 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 8

Decision-based evaluation for Task 2.

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	l_{TP}	$speed$	lw_{F1}
UNSL [26]	2	0.752	0.854	0.800	0.048	0.013	14.0	0.949	0.759
UNSL	0	0.752	0.767	0.760	0.048	0.017	15.0	0.945	0.718
UNSL	1	0.79	0.806	0.798	0.048	0.014	13.0	0.953	0.761
ELiRF-UPV [35]	0	**1.000**	0.883	**0.938**	**0.026**	**0.010**	4.0	0.988	**0.927**
Xabi_EHU [36]	0	0.846	0.961	0.900	0.030	0.012	8.0	0.973	0.875
Xabi_EHU	1	0.89	0.864	0.877	0.035	0.017	12.0	0.957	0.839
Xabi_EHU	2	0.79	0.913	0.847	0.036	0.015	13.0	0.953	0.807
Xabi_EHU	3	0.829	0.942	0.882	0.033	0.013	12.0	0.957	0.844
Xabi_EHU	4	0.756	0.961	0.846	0.031	0.013	8.0	0.973	0.823
OBSER-MENH [21]	0	0.048	**1.000**	0.092	0.064	0.049	3.0	0.992	0.092
OBSER-MENH	1	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091
OBSER-MENH	2	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091
OBSER-MENH	3	0.048	**1.000**	0.092	0.063	0.049	3.0	0.992	0.091
OBSER-MENH	4	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091
RELAI [25]	0	0.000	0.000	0.000	0.047	0.047			
RELAI	1	0.058	0.971	0.109	0.048	0.039	**1.0**	**1.000**	0.109
RELAI	2	0.058	0.971	0.109	0.048	0.039	**1.0**	**1.000**	0.109
RELAI	3	0.000	0.000	0.000	0.047	0.047			
RELAI	4	0.047	**1.000**	0.09	0.08	0.046	11.0	0.961	0.087
NLP-UNED-2 [37]	1	0.957	0.883	0.919	0.034	0.016	13.0	0.953	0.876
NLP-UNED-2	2	0.947	0.883	0.914	0.034	0.016	12.0	0.957	0.875
NLP-UNED-2	3	0.896	0.922	0.909	0.030	0.014	10.0	0.964	0.877
NLP-UNED-2	0	0.945	0.844	0.892	0.038	0.019	18.0	0.933	0.833
NLP-UNED-2	4	0.764	0.883	0.819	0.033	0.010	13.0	0.953	0.781
NUS-eRisk	4	0.062	0.951	0.117	0.059	0.040	6.0	0.981	0.114
NUS-eRisk	0	0.063	0.767	0.116	0.068	0.050	27.0	0.899	0.104
NUS-eRisk	1	0.06	0.903	0.113	0.068	0.043	13.0	0.953	0.107
NUS-eRisk	2	0.057	0.971	0.107	0.06	0.042	4.0	0.988	0.106
NUS-eRisk	3	0.067	0.874	0.125	0.065	0.042	17.0	0.938	0.117
BioNLP-IISERB [38]	0	0.933	0.68	0.787	0.038	0.037	62.0	0.766	0.603
BioNLP-IISERB	1	0.938	0.592	0.726	0.042	0.042	62.0	0.766	0.557
BioNLP-IISERB	3	**1.000**	0.049	0.093	0.045	0.045	**1.0**	**1.000**	0.093
BioNLP-IISERB	4	**1.000**	0.039	0.075	0.047	0.046	19.0	0.930	0.070
BioNLP-IISERB	2	0.000	0.000	0.000	0.047	0.047			
SINAI [39]	3	0.126	**1.000**	0.224	0.029	0.020	2.0	0.996	0.223
SINAI	0	0.115	**1.000**	0.206	0.029	0.021	**1.0**	**1.000**	0.206
SINAI	1	0.124	**1.000**	0.221	0.028	0.020	2.0	0.996	0.220
SINAI	2	0.108	**1.000**	0.195	0.03	0.022	**1.0**	**1.000**	0.195
SINAI	4	0.092	0.981	0.168	0.044	0.027	3.0	0.992	0.166
UMU [27]	1	**1.000**	0.388	0.559	0.047	0.043	94.5	0.651	0.364
UMU	0	0.086	**1.000**	0.158	0.039	0.029	2.0	0.996	0.157
UMU	2	0.048	**1.000**	0.092	0.057	0.044	2.0	0.996	0.091
UMU	3	0.593	0.311	0.408	0.048	0.045	80.0	0.701	0.286
UMU	4	0.048	**1.000**	0.091	0.053	0.045	2.0	0.996	0.090
NLP-UNED	0	0.057	0.903	0.108	0.052	0.052	**1.0**	**1.000**	0.108
NLP-UNED	1	0.053	0.845	0.099	0.064	0.063	141.0	0.502	0.050
NLP-UNED	2	0.054	0.854	0.101	0.056	0.055	**1.0**	**1.000**	0.101
NLP-UNED	3	0.055	0.874	0.103	0.056	0.056	**1.0**	**1.000**	0.103
NLP-UNED	4	0.066	0.728	0.121	0.071	0.071	142.0	0.500	0.060

Table 9 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 rounds. For tie breaking in the scores for the users, we used the traditional *docid* criteria (subject name). SINAI (run 1) obtained the best overall values after only one writing. At the other evaluation points,

Table 9
Ranking-based evaluation for Task 2.

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
UNSL [26]	0	**1.00**	**1.00**	0.46	**1.00**	**1.00**	0.70	**1.00**	**1.00**	0.64	**1.00**	**1.00**	0.64
UNSL	1	**1.00**	**1.00**	0.57	**1.00**	**1.00**	0.78	**1.00**	**1.00**	0.67	**1.00**	**1.00**	0.70
UNSL	2	**1.00**	**1.00**	0.55	**1.00**	**1.00**	0.75	**1.00**	**1.00**	0.69	**1.00**	**1.00**	0.69
ELiRF-UPV [35]	0	**1.00**	**1.00**	0.59	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	**0.95**	**1.00**	**1.00**	**0.94**
Xabi_EHU [36]	0	**1.00**	**1.00**	0.57	**1.00**	**1.00**	0.50	0.80	0.88	0.41	0.90	0.94	0.41
Xabi_EHU	1	**1.00**	**1.00**	0.56	0.90	0.94	0.49	0.70	0.76	0.38	0.80	0.88	0.40
Xabi_EHU	2	**1.00**	**1.00**	0.55	**1.00**	**1.00**	0.51	0.70	0.79	0.40	0.80	0.88	0.41
Xabi_EHU	3	**1.00**	**1.00**	0.56	0.90	0.94	0.51	0.80	0.86	0.41	0.90	0.94	0.42
Xabi_EHU	4	**1.00**	**1.00**	0.58	**1.00**	**1.00**	0.50	0.80	0.86	0.40	0.90	0.94	0.41
OBSER-MENH [21]	0	**1.00**	**1.00**	0.64	**1.00**	**1.00**	0.55	**1.00**	**1.00**	0.48	**1.00**	**1.00**	0.50
OBSER-MENH	1	**1.00**	**1.00**	0.65	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.49	**1.00**	**1.00**	0.50
OBSER-MENH	2	**1.00**	**1.00**	0.65	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.49	**1.00**	**1.00**	0.50
OBSER-MENH	3	**1.00**	**1.00**	0.64	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.48	**1.00**	**1.00**	0.50
OBSER-MENH	4	**1.00**	**1.00**	0.65	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.49	**1.00**	**1.00**	0.50
RELAI [25]	0	0.30	0.25	0.08	0.00	0.00	0.02	0.00	0.00	0.06			
RELAI	1	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00			
RELAI	2	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.02			
RELAI	3	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00			
RELAI	4	0.10	0.06	0.02	0.00	0.00	0.04	0.00	0.00	0.02			
NLP-UNED-2 [37]	0	**1.00**	**1.00**	0.32	**1.00**	**1.00**	0.83	**1.00**	**1.00**	0.90	**1.00**	**1.00**	0.90
NLP-UNED-2	1	**1.00**	**1.00**	0.40	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.93	**1.00**	**1.00**	**0.94**
NLP-UNED-2	2	**1.00**	**1.00**	0.40	**1.00**	**1.00**	0.85	**1.00**	**1.00**	0.92	**1.00**	**1.00**	0.93
NLP-UNED-2	3	**1.00**	**1.00**	0.59	**1.00**	**1.00**	0.92	**1.00**	**1.00**	0.95	**1.00**	**1.00**	0.93
NLP-UNED-2	4	**1.00**	**1.00**	0.57	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.87
NUS-eRisk	0	0.00	0.00	0.06	0.00	0.00	0.02	0.10	0.19	0.07	0.00	0.00	0.02
NUS-eRisk	1	0.00	0.00	0.03	0.00	0.00	0.02	0.10	0.19	0.06	0.00	0.00	0.01
NUS-eRisk	2	0.00	0.00	0.04	0.00	0.00	0.03	0.10	0.10	0.03	0.00	0.00	0.01
NUS-eRisk	3	0.00	0.00	0.06	0.00	0.00	0.04	0.10	0.19	0.08	0.00	0.00	0.02
NUS-eRisk	4	0.00	0.00	0.06	0.00	0.00	0.03	0.10	0.12	0.05	0.00	0.00	0.02
BioNLP-IISERB [38]	0	0.40	0.60	0.14									
BioNLP-IISERB	1	0.00	0.00	0.02									
BioNLP-IISERB	2	0.00	0.00	0.03									
BioNLP-IISERB	3	0.00	0.00	0.05									
BioNLP-IISERB	4	0.10	0.10	0.10									
SINAI [39]	0	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.85			
SINAI	1	**1.00**	**1.00**	**0.73**	**1.00**	**1.00**	0.90	**1.00**	**1.00**				
SINAI	2	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.84			
SINAI	3	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.86			
SINAI	4	0.80	0.86	0.53	0.90	0.94	0.56	0.70	0.80	0.47			
UMU [27]	0	**1.00**	**1.00**	0.63	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
UMU	1	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
UMU	2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.09	0.20	0.16	0.14
UMU	3	0.00	0.00	0.03	0.30	0.31	0.12	0.40	0.36	0.20	0.50	0.50	0.23
UMU	4	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.09	0.20	0.16	0.14
NLP-UNED	0	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02
NLP-UNED	1	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02
NLP-UNED	2	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02
NLP-UNED	3	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02
NLP-UNED	4	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02

ELiRF-UPV was again the best performing.

4. Task 3: Measuring the Severity of Eating Disorders

This task aims to estimate the severity of different symptoms associated with the diagnosis of eating disorders. Participants were provided with a thread of user submissions to analyze. For each user, we provided the participants with a collection of posts and comments, and participants were tasked with estimating the user's responses to a standardized eating disorder questionnaire based on the evidence found in the history of posts and comments.

The questionnaire used in this task is derived from the Eating Disorder Examination Questionnaire (EDE-Q)[6], which is a self-reported questionnaire comprising 28 items. It is adapted from the semi-structured interview Eating Disorder Examination (EDE)[7] [40]. For this task, our focus was on questions 1-12 and 19-28 from the EDE-Q. This questionnaire is specifically designed to assess various aspects and the severity of features associated with eating disorders. It consists of four subscales: Restraint, Eating Concern, Shape Concern, and Weight Concern, along with a global score. An excerpt of the EDE-Q is provided in Table 10.

Table 10: Excerpt of the 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
```

[6] https://www.corc.uk.net/media/1273/ede-q_quesionnaire.pdf

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

Table 10: Eating Disorder Examination Questionarie (continued)

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

⋮

22. Has your weight 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)

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)

Table 10: Eating Disorder Examination Questionarie (continued)

```
3. MODERATELY (3)
4. MODERATELY (4)
5. MARKEDLY (5)
6. MARKEDLY (6)
```

The main objective of this task was to investigate the feasibility of automatically estimating the severity of multiple symptoms associated with eating disorders. The participating algorithms estimated the user's response to each question based on their writing history. To evaluate the performance of the systems, we collected questionnaires completed by users on social media, along with their corresponding writing history. These user-completed questionnaires served as the ground truth against which the responses provided by the systems were evaluated.

During the training phase, participants were provided with data from 28 users who participated in the 2022 edition [14]. This training data included the writing history of the users as well as their responses to the EDE-Q questions. In the test phase, there were 46 new users for whom the participating systems had to generate results. The expected format for submitting the results followed a specific file format:

```
username1 answer1 answer2...answer12 answer19...answer28
username2 answer1 answer2...answer12 answer19...answer28
⋮
```

Each line has the username and 22 values (no answers from 13 to 18). These values correspond with the responses to the questions above (the possible values are 0,1,2,3,4,5,6).

4.1. 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 [41]).

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

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).

- **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.

- **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) [40]. 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. Results

Table 11 presents the results obtained by the participants in this task. To provide context, the table includes the performance of three baseline variants in the top block: "all 0s," "all 6s," and "average." The "all 0s" variant represents a strategy where the same response (0) is submitted for all questions. Similarly, the "all 6s" variant submits response 6 for all questions. The "average" variant calculates the mean of the responses provided by all participants for each question and submits the response that is closest to this mean value (e.g., if the mean response provided by the participants is 3.7, then this average approach would submit a 4).

The results indicate that the top-performing system in terms of Mean Absolute Error (MAE) was run 0 by UMU. However, this particular run did not outperform the naive baseline approach of submitting all 0s in terms of Mean Zero-One Error (MZOE). Among the participating systems, GMU-FAST' run 3 achieved the best performance in two metrics: MAE_{macro} and Global ED. For the Eating Concern Subscale, the best-performing system was GMU-FAST' run 1, while for the Shape subscale, RiskBusters' run 4 showed the highest performance.

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 2023 working notes proceedings for the participants.

BFH-AMI [42]. The BFH-AMI team, affiliated with Applied Machine Intelligence of Switzerland, participated in Task 3 of the eRisk 2023 challenge. The team employed a logistic regression model that incorporated user and question embeddings derived from the Large Language Models. To

Table 11
Task 3 Results. Participating teams and runs with corresponding scores for the metrics.

team	run ID	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
baseline	all0s	2.419	**0.674**	2.803	3.207	2.138	3.221	3.028	2.682
baseline	all6s	3.581	0.834	3.995	3.839	4.814	3.650	3.950	3.318
baseline	average	2.091	0.859	1.957	2.391	**1.592**	2.398	2.162	**2.002**
BFH-AMI [42]	0	2.407	0.719	2.729	3.169	2.597	2.854	2.923	2.414
GMU-FAST	0	2.529	0.902	2.012	2.498	2.585	1.948	1.950	2.221
GMU-FAST	1	2.525	0.903	1.992	2.487	2.584	**1.924**	1.917	2.219
GMU-FAST	2	2.738	0.764	2.058	2.708	2.278	2.641	2.295	2.662
GMU-FAST	3	2.671	0.833	**1.741**	**1.999**	2.740	2.053	2.083	2.401
GMU-FAST	4	2.534	0.796	1.879	2.174	2.469	2.136	2.033	2.387
RiskBusters [43]	0	2.338	0.691	1.922	2.294	1.866	2.492	1.999	2.425
RiskBusters	1	2.352	0.699	1.858	2.127	2.025	2.365	2.034	2.466
RiskBusters	2	2.396	0.704	1.861	2.178	1.859	2.484	1.957	2.468
RiskBusters	3	2.419	0.709	1.898	2.251	1.935	2.440	2.037	2.445
RiskBusters	4	2.346	0.705	1.859	2.217	1.862	2.398	**1.898**	2.395
RiskBusters	5	2.334	0.702	1.854	2.230	1.898	2.381	1.947	2.378
RiskBusters	6	2.408	0.696	1.936	2.365	2.048	2.536	1.985	2.414
RiskBusters	7	2.347	0.696	1.975	2.534	1.911	2.443	2.215	2.494
UMU [27]	0	**2.194**	0.800	2.027	2.288	1.777	2.412	2.556	2.135

address the challenge, the authors employed two different models, namely BERT and GPT-Large, to generate embeddings for users' posts and publications. These embeddings were then utilized as input features for the logistic regression model. The aim was to capture the underlying patterns and information within the social media writing history of patients that could help predict their responses to the EDE-Q. The use of logistic regression allowed the team to build a predictive model that could estimate the likelihood of different responses based on the extracted features. By incorporating user and question embeddings the model could effectively leverage the semantic information and context within the social media writings.

BLUE [23]. The BLUE team participated in Task 1 of the eRisk 2023 challenge. Their approach involved using dense retrievers for semantic search, with the aim of retrieving relevant sentences from a collection. They employed two types of queries for the search process. First, they utilized the descriptions of each symptom found in the BDI-II questionnaire. These symptom descriptions served as queries for the search. Additionally, they leveraged ChatGPT, a language model, to generate synthetic queries for each symptom in the BDI-II questionnaire. The team believed that using ChatGPT to generate synthetic queries would introduce more diversity in expressions and potentially yield more relevant sentences during the search process. The team conducted five runs of their approach and employed two transformer-based models for embedding the social media posts. These models were the original and generated responses of the BDI-II, MentalRoBERTa, and a variant of MPNet. Surprisingly, the model that performed semantic

Table 12
eRisk 2023 participants.

team	Task 1 #runs	Task 2 #runs	Task 3 #runs
[42] BFH-AMI			1
[23] BLUE	5		
[38] BioNLP-IISERB		5	
[35] ELiRF-UPV		1	
[20] Formula-ML	4		
GMU-FAST	2		5
[28] Mason-NLP	1		
[24] NailP	5		
NLP-UNED		5	
[37] NLP-UNED-2		5	
NUS-eRisk		5	
[21] OBSER-MENH	5	5	
[25] RELAI	5	5	
[43] RiskBusters			8
[39] SINAI		5	
[27] UMU	2	5	1
[26] UNSL	3	3	
[22] uOttawa	5		
[36] Xabi_EHU		5	

search using the initial responses from the BDI-II questionnaire as queries achieved better performance compared to the model using generated queries. Furthermore, it was observed that the generated synthetic data was too specific for this particular task.

BioNLP-IISERB [38]. BioNLP-IISERB participated in Task 2 of the challenge. The team submitted five different runs, employing various text-mining frameworks and strategies for text classification and feature engineering. To extract features from the social media posts, BioNLP-IISERB utilized both traditional and transformer-based approaches. They employed the bag of words model, specifically TF-IDF weighting, as well as transformer-based embeddings such as BERT, Longformer, and RoBERTa. For the classification stage, the team explored several classifiers to determine their effectiveness in the task. They utilized classifiers such as Adaptive Boosting (AdaBoost), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and transformer-based classifiers. Among their different approaches, BioNLP-IISERB found that combining a Support Vector Machine and an Adaptive Boosting classifier yielded their most successful performance.

ELiRF-UPV [35]. ELiRF-UPV, associated with the Valencian Research Institute for Artificial Intelligence and the informatics School of Universitat Politècnica de València, participated in the eRisk 2023 Task 2. Their approach centered around utilizing SVMs, a traditional machine learning technique. The ELiRF-UPV team employed SVMs with different kernels and regularization parameters in their experiments. They utilized cross-validation techniques to assess the performance of the SVM classifiers. To represent the text data, they employed a TF-IDF vectorizer with a maximum limit of 5,000 features. Notably, the ELiRF-UPV team achieved the highest scores in Precision, F1 score, ERDE5, ERDE50, and latency-weighted F1 in the eRisk 2023 Task 2.

Formula-ML [20]. The Formula-ML team participated in Task 1. To represent the sentences, Formula-ML utilized various techniques such as word2vec, the TF-IDF model, and sentence transformers. These techniques allowed them to convert the text data into numerical representations or embeddings. To determine the degree of correlation between the sentences and the answers from the BDI questionnaire, Formula-ML utilized a soft cosine similarity measure. This measure assessed the relatedness of the topics discussed in the sentences to the symptoms of depression. The results obtained by Formula-ML varied depending on the embedding model used. The models based on SentenceTransformers achieved great performance and even ranked highly in some evaluation metrics. However, the word2vec-based models did not reach the same level of performance.

Mason-NLP [28]. The Mason-NLP team from George Mason University participated in Task 1 of the CLEF eRisk lab. Their approach incorporated two deep learning models, MentalBERT and RoBERTa, along with LSTM. The team employed several steps to reduce the number of sentences before passing them through the MentalBERT model, which computed an embedding representation for each sentence. The ranking of each symptom was then determined by calculating the cosine similarity between the embedding representation of each sentence and the embedding representation of the symptom options. However, the evaluation results of their approach did not meet the expectations of the authors. Despite this outcome, they acknowledged the opportunity for improvement and proposed several avenues for future work. One suggestion was to fine-tune MentalBERT using BDI training data, which could potentially enhance the model's performance.

NailP [24]. The NailP team participated in Task 1. Their approach focused on data pre-processing and calculating similarity between text representations. Initially, the authors performed data pre-processing by selecting publications that contained self-referential content, specifically highlighting personal pronouns. They then filtered out positive or neutral sentences, focusing only on those with potentially negative sentiment. Sentence embeddings were calculated using SBERT (Sentence-BERT), a model that generates contextualized sentence embeddings. Additionally, the team obtained embeddings from the descriptions of the BDI-II items using the same SBERT model. To rank the sentences, a semantic search was conducted with different filters to include or exclude negative posts.

NLP-UNED-2 [37]. The NLP-UNED-2, from the Spanish National University of Distance Education, participated in Task 2 of the eRisk 2023 challenge. Their approach consisted of dataset relabeling using Approximate Nearest Neighbors (ANN) on vectorial representations of messages. ANN is a technique that finds approximate matches for a given query in a high-dimensional space. By utilizing ANN, the team aimed to improve the labeling of the dataset. For classification, the UNED team employed neural networks. These neural networks were used to classify and identify instances of pathological gambling within the social media data.

OBSER-MENH [21]. The OBSER-MENH team participated in both Task 1 and Task 2 of the eRisk challenge. For Task 1, their approach involved using SBERT to compute vector representations of each publication. They then utilized the descriptions of the BDI-II as queries to rank depressive symptoms. By leveraging SBERT and the BDI-II descriptions, they aimed to assess the relevance of each symptom in the publications. In Task 2, the OBSER-MENH team focused on addressing class imbalance and preventing overfitting. They employed an ensemble approach that combined variants of three models. The ensemble aimed to analyze the different penalty weights applied to a feedforward neural network (FNN). This approach allowed them to train multiple models with varying penalty weights, helping to find an optimal balance and prevent overfitting.

RELAI [25]. The RELAI team, a collaboration between universities in Quebec and McMaster, participated in both Task 1 and Task 2. For Task 1, the team utilized transformer-based sentence encoders and employed the Okapi BM25 method to select sentences. The selected sentences were then subjected to a similarity task, and the similarity scores were used for ranking, with different strategies employed. While their results did not yield significant improvements, they did outperform their chosen simple baseline in some cases. In Task 2, the RELAI team employed "lightweight" approaches. One approach involved extracting stylometric and shallow features, such as character and n-gram frequencies and sentence length. These features were then fed into a multilayer perceptron for classification. Another approach utilized topic modeling of the posts, with the resulting topics used as input to a multilayer perceptron. The final approach took a probabilistic approach, estimating the proportion of positive users from past editions to determine the gamma distribution from which they emerge.

RiskBusters [43]. The RiskBuster team from the University of Bucharest participated in Task 3. Their approach involved the use of a transformer-based topic modeling method to measure the severity of eating disorder symptoms. The team customized the BERTopic framework to obtain topic distributions at the user level. These topic distributions were then utilized as input features for downstream classification tasks. By leveraging topic modeling, the team aimed to capture the underlying themes and severity of eating disorder symptoms within the social media writings. To improve the quality of embeddings, the RiskBuster team adapted MentalBERT, a transformer-based language model, to the eating disorder domain.

SINAI [39]. The SINAI team, a collaboration between Jaén and Bocconi universities, participated Task 2. They developed various approaches using XLM-RoBERTa and RoBERTa transformer models, with their most innovative proposal combining LSTM and RoBERTa architectures. Their approach involved encoding each user's post using RoBERTa to obtain embeddings, which were then passed through an LSTM. The final prediction was made using a Feed Forward Network. The team also designed pre-processing steps for the corpus before training. With this approach, they achieved the highest Recall score among all participants for the binary task and had the second-highest ERDE value. In terms of ranking-based evaluation, they achieved one of the highest positions.

UMU [27]. The UMU team introduced different strategies to address the three distinct tasks in the eRisk 2023 challenge. For Task 1, the team approached it as a question-answering problem and presented two proposals: one for monolingual text and another for multilingual text. They employed a pre-trained sentence transformer model to assess the severity of each depressive symptom within users' text collections. The team used a depression-domain lexicon approach to select texts closely related to depression and ranked the 21 depression symptoms of the BDI questionnaire. By treating Task 1 as a question-answering task, they linked the questions of the BDI questionnaire to their possible answers in users' writings. This approach allowed them to determine the degree of relationship between each text and the 21 symptoms of depression. For Task 2, which focused on the early detection of signs of pathological gambling, UMU team presented five runs that combined decision-making strategies with classification models. In Task 3, UMU team presented a run based on fine-tuning a pre-trained sentence transformer model. They processed the dataset and performed emoji feature extraction to enhance the model's performance.

UNSL [26]. The UNSL (Universidad Nacional de San Luis) team participated in both Task 1 and Task 2. For Task 1, they submitted three proposals. The first two proposals focused on the similarity of contextualized embedding vectors between the posts and the symptoms of the Beck Depression Inventory. The third proposal utilized a prompting strategy, reformulating the sentence retrieval task into a masked language problem. They used ChatGPT to generate synthetic examples related to a specific symptom, and a RoBERTa model was fine-tuned to solve the masking language problem. The prompting-based strategy yielded better results than the similarity-based proposals, providing promising outcomes for information retrieval. For Task 2, related to the early detection of signs of pathological gambling, the UNSL team proposed three fine-tuned models, followed by a decision policy based on criteria defined by an early detection framework. One of the models incorporated an extended vocabulary with important words specific to the domain. The UNSL models demonstrated good performances in decision-based metrics, ranking-based metrics, and runtime for Task 2.

uOttawa [22]. The uOttawa team, representing the University of Ottawa, participated in Task 1. They employed a combination of word embedding models, including GloVe, DistilBERT, RoBERTa, and the Universal Sentence Encoder (USE). Among the models submitted by the uOttawa team, the Universal Sentence Encoder (USE) consistently performed the best for every metric, whether using majority voting or unanimity for the qrels. The methods based on contextual representations, including USE, RoBERTa, and DistilBERT, outperformed the methods based on GloVe embeddings. After empirical analysis, the team hypothesized that the better performance of contextual-based methods could be attributed to the removal of pronouns, which may contain relevant information in the discourse. This removal was not applied to the methods based on GloVe embeddings.

Xabi_EHU [36]. The Xabi_EHU team, representing the University of the Basque Country (UPV/EHU), participated in Task 2. To address the issue of class imbalance, where the Pathological Gamblers group is smaller than the Control group, the team employed a neural network with a customized loss function. This approach allowed them to adjust penalties for false positives and false negatives, providing flexibility to meet specific requirements. By utilizing this customized loss function, their training approach did not penalize false positives and negatives equally.

6. Conclusions

This paper provides an overview of eRisk 2023, which marked the seventh edition of the lab. The lab focused on three distinct tasks: symptoms search (Task 1 on depression), early detection (Task 2 on pathological gambling), and severity estimations (Task 3 on eating disorders). In Task 1, participants were presented with a collection of sentences and tasked with ranking them based on their relevance to each symptom of depression outlined in the BDI-II questionnaire. Task 2 required participants to sequentially analyze social media posts and issue alerts for individuals displaying signs of gambling risk. In Task 3, participants were provided with the complete user

history and required to automatically estimate the user's responses to a standardized depression questionnaire.
A total of 105 runs were submitted by 20 teams for the proposed tasks. While the effectiveness of the solutions varies across tasks, the experimental results highlight the value of extracting evidence from social media. This suggests the promising potential of automatic or semi-automatic screening tools for detecting at-risk individuals. These findings emphasize the necessity for the development of benchmarks for text-based risk indicator screening.

Acknowledgments

This work was supported by project 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). The first and second authors thank the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (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 thanks the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) 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, 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, 2022, p. 233–256.

[15] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2022: Early risk prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5–8, 2022, 2022, pp. 821–850.

[16] 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.

[17] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early

risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, 2023.

[18] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2023: Depression, pathological gambling, and eating disorder challenges, in: Advances in Information Retrieval - 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, 2023, p. 585–592.

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

[20] N. R, P. Bolimera, Y. Gupta, A. K. M, Exploring depression symptoms through similarity methods in social media posts, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[21] J. Martinez-Romo, L. Araujo, X. Larrayoz, M. Oronoz, A. Pérez, OBSER-MENH at eRisk 2023: Deep learning-based approaches for symptom detection in depression and early identification of pathological gambling indicators, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[22] Y. Wang, D. Inkpen, uOttawa at eRisk 2023: Search for symptoms of depression, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[23] A.-M. Bucur, Utilizing ChatGPT generated data to retrieve depression symptoms from social media, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[24] E. Bezerra, L. dos Santos, R. Nascimento, R. P. Lopes, G. Guedes, NailP at eRisk 2023: Search for symptoms of depression, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[25] D. Maupomé, T. Soulas, F. Rancourt, G. Cantin-Savoie, G. Winterstein, S. Mosser, M.-J. Meurs, Lightweight methods for early risk detection, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[26] H. Thompson, L. Cagnina, M. Errecalde, Strategies to harness the transformers' potential: UNSL at eRisk 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[27] R. Pan, J. A. G. Díaz, R. Valencia-Garcia, UMUTeam at eRisk@CLEF 2023 shared task: Transformer models for early detection of pathological gambling, depression, and eating disorder, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[28] F. A. Sakib, A. A. Choudhury, Özlem Uzuner, MASON-NLP at eRisk 2023: Deep learning-based detection of depression symptoms from social media texts, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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).

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

[35] A. M. Marco, X. Huang, L.-F. Hurtado, F. Pla, ELiRF-UPV at eRisk 2023: Early detection of pathological gambling using SVM, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[36] X. Larrayoz, N. Lebeña, A. Casillas, A. Pérez, Representation exploration and deep learning applied to the early detection of pathological gambling risks, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[37] H. Fabregat, A. Duque, L. Araujo, J. Martinez-Romo, NLP-UNED-2 at eRisk 2023: Detecting pathological gambling in social media through dataset relabeling and neural networks, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[38] A. Talha, T. Basu, A natural language processing based risk prediction framework for pathological gambling, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[39] A. M. Mármol-Romero, F. M. P. del Arco, A. Montejo-Ráez, Nsinai at erisk@clef 2023: Approaching early detection of gambling with natural language processing, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[40] C. G. Fairburn, Z. Cooper, M. O'Connor, Eating disorder examination Edition 17.0D (April, 2014).

[41] S. Baccianella, A. Esuli, F. Sebastiani, Evaluation measures for ordinal regression, 2009, pp. 283–287. doi:10.1109/ISDA.2009.230.

[42] G. Merhbene, A. R. Puttick, M. Kurpicz-Briki, BFH-AMI at eRisk@CLEF 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.

[43] D.-N. Grigore, I. Pintilie, Transformer-based topic modeling to measure the severity of eating disorder symptoms, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18-21, 2023.