Overview of eRisk at CLEF 2024: Early Risk Prediction on the Internet (Extended Overview)

Notebook for the eRisk Lab at CLEF 2024

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 presents eRisk 2024, the eighth edition of the CLEF conference's lab dedicated to early risk detection. Since its inception, the lab has been at the forefront of developing and refining evaluation methodologies, effectiveness metrics, and processes for early risk detection across various domains. These early alerting models hold significant value, particularly in sectors focused on health and safety, where timely intervention can be crucial. eRisk 2024 featured three main tasks designed to push the boundaries of early risk detection techniques. The first task challenged participants to rank sentences based on their relevance to standardized depression symptoms, a crucial step in identifying early signs of depression from textual data. The second task focused on the early detection of anorexia indicators, aiming to develop models that can recognize the subtle cues of this eating disorder before it becomes critical. The third task was centered around estimating responses to an eating disorders questionnaire by analyzing users' social media posts. Participants had to leverage the rich, real-world textual data available on social media to gauge potential mental health risks. Through these tasks, eRisk 2024 continues to advance the field of early risk detection, fostering innovations that could lead to significant improvements in public health interventions.

Keywords

Early risk, Depression, Anorexia, Eating disorders

1. Introduction

The primary goal of eRisk is to explore evaluation methodologies, metrics, and other factors essential for developing research collections and identifying early risk signs. Early detection technologies are increasingly important in safety and health fields. These technologies are particularly useful for detecting mental illness symptoms, identifying interactions between infants and sexual abusers, or spotting antisocial threats online, where they can provide early warnings and potentially prevent harmful outcomes.

Our lab focuses on a range of psychological issues, including depression, self-harm, pathological gambling, and eating disorders. We have found that the relationship between psychological conditions and language use is complex, highlighting the need for more effective automatic language-based screening models. This complexity arises from the subtle and varied ways in which psychological distress can manifest in language, necessitating sophisticated analytical techniques.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*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)

In 2017, we initiated our efforts with a task aimed at detecting early signs of depression. This task utilized new evaluation methods and a test dataset described in [1, 2]. The goal was to develop models capable of identifying depressive symptoms from textual data, which could then be used for early intervention. In 2018, we expanded our scope to include the early detection of anorexia [3, 4]. This task required models to identify language patterns indicative of anorexia, providing a tool for early diagnosis.

In 2019, we continued our work on anorexia and introduced new challenges. These included detecting early signs of self-harm and estimating responses to a depression questionnaire based on social media activity [5, 6, 7]. The self-harm detection task aimed to identify individuals at risk by analyzing their online posts for signs of self-injurious behavior. The severity estimation task aimed to quantify the level of depressive symptoms exhibited in social media posts, providing a more nuanced understanding of an individual's mental health status. In 2020, our focus included further development of self-harm detection and a new task on depression severity estimation [8, 9, 10].

In 2021, we concentrated on early detection tasks for pathological gambling and self-harm, along with a task for estimating depression severity [11, 12, 13]. The pathological gambling task involved identifying language patterns associated with gambling addiction, which could be used to flag individuals at risk. The self-harm and depression severity tasks continued to build on our previous work, refining the models and evaluation methods.

The 2022 edition of eRisk introduced tasks for early detection of pathological gambling, depression, and severity estimation of eating disorders [14, 15, 16]. These tasks aimed to improve the accuracy and reliability of early detection models, providing valuable tools for mental health professionals.

In 2023, eRisk tasks included ranking sentences by their relevance to depression symptoms, early detection of gambling signs, and severity estimation of eating disorders [17, 18, 19]. The sentence ranking task required models to assess the relevance of individual sentences to specific depressive symptoms, as outlined in the BDI-II questionnaire. This task aimed to enhance the precision of symptom identification in textual data.

In 2024, eRisk presented three campaign-style tasks [20, 19]. The first task focused on ranking sentences related to the 21 symptoms of depression as per the BDI-II questionnaire, using sentences extracted from social media posts. The second task continued our work on early detection of anorexia, requiring models to identify language patterns indicative of this eating disorder. The third task revisited the severity estimation of eating disorders, aiming to quantify the severity of symptoms exhibited in textual data. Detailed descriptions of these tasks are provided in the subsequent sections of this overview article.

In 2024, we had 84 teams registered for the lab. We received results from 17 of them: 29 runs for Task 1, 44 runs for Task 2, and 14 for Task 3. These results provided valuable insights into the effectiveness of different models and approaches, contributing to the ongoing development of early detection technologies.

2. Task 1: Search for Symptoms of Depression

This task builds on eRisk 2023's Task 1, which focused on ranking sentences from user writings based on their relevance to specific depression symptoms. Participants had to order sentences according to their relevance to the 21 standardized symptoms listed in the BDI-II Questionnaire [21]. A sentence was considered relevant if it reflected the user's condition related to a symptom, including positive statements (e.g., "I feel quite happy lately" is relevant for the symptom "Sadness"). This year, the dataset included the target sentence and the sentences immediately before and after it to provide additional context.

2.1. Dataset

The dataset provided was in TREC format, tagged with sentences derived from eRisk's historical data. Table 1 presents some statistics of the corpus.

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

Number of users	551,311
Number of sentences	15,542,200
Average number of words per sentence	17.98

```
1Q0251001_0_1000110myGroupNameMyMethodName
1Q0251202_5_400029.5myGroupNameMyMethodName
1Q0858202_3_200039myGroupNameMyMethodName
...
21Q0153202_2_209981.25myGroupNameMyMethodName
21Q0331302_1_109991myGroupNameMyMethodName
21Q0223133_9_810000.9myGroupNameMyMethodName
```

Figure 1: Example of a participant's run.

2.2. Assessment Process

Participants were given the corpus of sentences and the description of the symptoms from the BDI-II questionnaire. They were free to decide on the best strategy to derive queries for representing the BDI-II symptoms. Each team could submit up to 5 variants (runs). Each run included 21 TREC-style formatted rankings of sentences, as shown in Figure 1. For each symptom, participants submitted up to 1000 results sorted by estimated relevance. We received 29 runs from 9 participating teams (see Table 2).

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

Team	# of submissions
ThinkIR	1
SINAI [22]	2
RELAI [23]	5
NUS-IDS [24]	5
MindwaveML [25]	3
MeVer-REBECCA [26]	2
GVIS	1
DSGT [27]	5
APB-UC3M [28]	5
Total	29

To create the relevance judgments, three assessors annotated a pool of sentences associated with each symptom. These candidate sentences were obtained by performing top-k pooling from the relevance rankings submitted by the participants.

The assessors were given specific instructions (see Figure 2) dto determine the relevance of candidate sentences. They considered a sentence relevant if it addressed the topic and provided explicit information about the individual's state in relation to the symptom. This dual concept of relevance (on-topic and reflective of the user's state with respect to the symptom) introduced a higher level of complexity compared to standard relevance assessments. Consequently, we developed a robust annotation methodology and formal assessment guidelines to ensure consistency and accuracy. The main change from eRisk 2023's assessment process was that the assessors were presented with the sentence and its context (previous and following sentences, if available).

To create the pool of sentences for assessment, we implemented top-k pooling with $k = 50$. The resulting pool sizes per sentence are reported in Table 3.

The annotation process involved a team of three assessors with different backgrounds and expertise.

```
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 individual
     related to the topic of the BDI item. But it is not necessary that the exact same words
     are used. Assessors should label the relevance of the sentence taking into account its
     context (preceding and following sentence).

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). Assessors should label the relevance of the
     sentence taking into account its context (preceding and following sentence). 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).

One assessor had professional training in psychology, while the other two were computer science researchers—a postdoctoral fellow and a Ph.D. student—with a specialization in early risk technologies. To ensure consistency and clarity throughout the process, the lab organizers conducted a preparatory session with the assessors. During this session, an initial version of the guidelines was discussed, and any doubts or questions raised by the assessors were addressed. This collaborative effort resulted in the final version of the guidelines[1].

According to these guidelines, a sentence is considered 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 basis for determining the relevance of sentences during the annotation process.

The final outcomes of the annotation process are presented in Table 3, where the number of relevant sentences per BDI item is reported (last two columns). We marked a sentence as relevant following two aggregation criteria: unanimity and majority.

[1] https://erisk.irlab.org/guidelines_erisk24_task1.html

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

BDI Item (#)	pool	# rels (3/3)	# rels (2/3)
Sadness (1)	783	226	442
Pessimism (2)	747	122	294
Past Failure (3)	715	160	270
Loss of Pleasure (4)	652	116	196
Guilty Feelings (5)	737	311	399
Punishment Feelings (6)	611	87	162
Self-Dislike (7)	730	308	385
Self-Criticalness (8)	700	187	281
Suicidal Thoughts or Wishes (9)	701	326	410
Crying (10)	755	311	433
Agitation (11)	758	276	400
Loss of Interest (12)	657	131	211
Indecisiveness (13)	784	164	308
Worthlessness (14)	567	222	258
Loss of Energy (15)	609	181	243
Changes in Sleeping Pattern (16)	777	244	365
Irritability (17)	727	192	305
Changes in Appetite (18)	694	219	334
Concentration Difficulty (19)	581	204	286
Tiredness or Fatigue (20)	682	238	343
Loss of Interest in Sex (21)	847	137	304

2.3. Results

The performance results for the participating systems are shown in Tables 4 (majority-based qrels) and 5 (unanimity-based qrels). The tables report several standard performance metrics, such as Mean Average Precision (MAP), mean R-Precision, mean Precision at 10, and mean NDCG at 1000. Run Config_5, from the team NUS-IDS [24], achieved the top-ranking performance for nearly all metrics and relevance judgment types. It consists in an ensemble model designed for computing semantic similarity with respect to different expanded descriptions of BDI symptoms. This ensemble leverages three pre-trained language models: `all-mpnet-base-v2`, `all-MiniLM-L12-v2`, and `all-distilroberta-v1`.This approach is similar to the APB-UC3M [28] team's proposal, which achieved the best results in terms of P@10 using majority voting. In contrast, the MeVer-REBECCA [26] team opted for the `bge-small-en-v1.54` embedding model, attaining the highest P@10 scores in the unanimity case.

3. Task 2: Early Detection of Signs of Anorexia

This task is the third edition of the challenge to develop models for early identification of anorexia signs. The goal was to process evidence sequentially and detect early indications of anorexia as soon as possible. Participating systems analyzed user posts on social media in the order they were written. Successful outcomes from this task could be used for sequential monitoring of user interactions across various online platforms like blogs, social networks, and other digital media.
The test collection used for this task followed the format described by Losada and Crestani [29]. It contains writings, including posts and comments, from a selected group of social media users. Users are categorized into two groups: anorexia and non-anorexia. For each user, the collection contains a sequence of writings arranged in chronological order. To facilitate the task and ensure uniform distribution, we established a server that systematically provided user writings to the participating teams. Further details about the server's setup are available on the lab's official website[2].
This was a train-test task. During the training stage, teams had access to the entire history of writings

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

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

Team	Run	AP	R-PREC	P@10	NDCG
ThinkIR	BM25Similarity	0.203	0.258	0.881	0.410
SINAI	SINAI_DR_majority_daug	0.064	0.107	0.562	0.174
SINAI	GPT3-Insight-8	0.008	0.024	0.200	0.044
RELAI	RELAI_paraphrase-MiniLM-L12-v2	0.267	0.346	0.738	0.525
RELAI	RELAI_paraphrase-MiniLM-L6-v2	0.236	0.325	0.590	0.503
RELAI	RELAI_all-MiniLM-L6-v2-simcse	0.226	0.322	0.595	0.495
RELAI	tfidf_sgd	0.163	0.240	0.552	0.394
RELAI	RELAI_word2vec	0.000	0.000	0.000	0.000
NUS-IDS	Config_5	**0.375**	**0.434**	0.924	**0.631**
NUS-IDS	Config_2	0.352	0.415	0.881	0.616
NUS-IDS	Config_4	0.336	0.401	0.890	0.599
NUS-IDS	Config_1	0.312	0.386	0.871	0.576
NUS-IDS	Config_3	0.286	0.359	0.857	0.556
MindwaveML	Mindwave-MLMiniLML12MLP_weighted	0.159	0.240	0.567	0.396
MindwaveML	Mindwave-MLMiniLML12MLP_0.5	0.149	0.231	0.538	0.378
MindwaveML	Mindwave-MLMiniLML12	0.133	0.212	0.490	0.335
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity_gpt	0.301	0.340	0.981	0.506
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity	0.295	0.332	0.976	0.517
GVIS	GVIS	0.000	0.002	0.035	0.005
DSGT	logistic_transformer_v5	0.000	0.009	0.000	0.014
DSGT	logistic_word2vec_v5	0.000	0.001	0.000	0.003
DSGT	count_logistic	0.000	0.000	0.000	0.001
DSGT	count_nb	0.000	0.000	0.000	0.000
DSGT	word2vec_logistic	0.000	0.000	0.000	0.000
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L6-v2	0.354	0.391	**0.986**	0.591
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L12-v2	0.337	0.378	0.990	0.564
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.293	0.330	0.967	0.525
APB-UC3M	APB-UC3M_ensemble	0.057	0.120	0.324	0.191
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.056	0.118	0.371	0.206

for training users. We indicated which users had explicitly mentioned being diagnosed with anorexia. Participants could tune their systems with this training data. In 2024, the training data included users from previous editions of the anorexia task (2018 and 2019).

During the test stage, participants connected to our server and engaged in an iterative process of receiving user writings and sending their responses. At any point within the chronology of user writings, participants could halt the process and issue an alert. After reading each user writing, teams had to decide between two options: i) alerting about the user, indicating a predicted sign of anorexia, or ii) not alerting about the user. Participants made this choice independently for each user in the test split. Once an alert was issued, it was final, and no further decisions regarding that individual were considered. Conversely, the absence of alerts was non-final, allowing participants to submit an alert later if they detected signs of risk.

We evaluated the systems' performance using two indicators: the accuracy of the decisions made and the number of user writings required to reach those decisions. These criteria provide insights into the effectiveness and efficiency of the systems. To support the test stage, we deployed a REST service. The server iteratively distributed user writings and waited for responses from participants. New user data was not provided to a participant until the service received a decision from that team. The submission period for the task was open from February 5th, 2024, until April 12th, 2024.

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

Team	Run	MAP	R-PREC	P@10	NDCG
ThinkIR	BM25Similarity	0.174	0.246	0.652	0.417
SINAI	SINAI_DR_majority_daug	0.046	0.098	0.362	0.150
SINAI	GPT3-Insight-8	0.001	0.009	0.052	0.014
RELAI	RELAI_paraphrase-MiniLM-L12-v2	0.248	0.329	0.576	0.537
RELAI	RELAI_paraphrase-MiniLM-L6-v2	0.207	0.287	0.410	0.509
RELAI	RELAI_all-MiniLM-L6-v2-simcse	0.194	0.275	0.433	0.499
RELAI	tfidf_sgd	0.138	0.207	0.376	0.383
RELAI	RELAI_word2vec	0.000	0.000	0.000	0.000
NUS-IDS	Config_5	**0.392**	**0.436**	0.795	**0.692**
NUS-IDS	Config_2	0.370	0.431	0.752	0.677
NUS-IDS	Config_4	0.358	0.416	0.771	0.662
NUS-IDS	Config_1	0.329	0.391	0.786	0.636
NUS-IDS	Config_3	0.312	0.375	0.757	0.621
MindwaveML	Mindwave-MLMiniLML12MLP_weighted	0.158	0.238	0.471	0.427
MindwaveML	Mindwave-MLMiniLML12MLP_0.5	0.147	0.227	0.457	0.408
MindwaveML	Mindwave-MLMiniLML12	0.128	0.203	0.410	0.360
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity_gpt	0.305	0.357	**0.833**	0.551
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity	0.294	0.349	0.824	0.556
GVIS	GVIS	0.000	0.002	0.030	0.004
DSGT	logistic_transformer_v5	0.000	0.006	0.000	0.010
DSGT	logistic_word2vec_v5	0.000	0.001	0.000	0.003
DSGT	count_logistic	0.000	0.000	0.000	0.000
DSGT	count_nb	0.000	0.000	0.000	0.000
DSGT	word2vec_logistic	0.000	0.000	0.000	0.000
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L6-v2	0.345	0.407	0.829	0.630
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L12-v2	0.333	0.389	0.805	0.608
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.285	0.342	0.776	0.561
APB-UC3M	APB-UC3M_ensemble	0.052	0.106	0.248	0.193
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.033	0.084	0.190	0.169

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

	Anorexia	*Control*
Num. subjects	92	692
Num. submissions (posts & comments)	28,043	338,843
Avg num. of submissions per subject	304.8	489.6
Avg num. of days from first to last submission	$\approx$ 482	$\approx$ 971
Avg num. words per submission	28.5	21.4

To construct the ground truth assessments, we adopted established approaches to optimize the use of assessors' time, as documented in previous studies [30, 31]. These methods employ simulated pooling strategies to create effective test collections. The main statistics of the test collection used for T2 are presented in Table 6.

3.1. Decision-based Evaluation

This evaluation approach uses the binary decisions made by the participating systems for each user. In addition to standard classification measures such as Precision, Recall, and F1 score (computed with respect to the positive class), we also calculate ERDE (Early Risk Detection Error), used in previous editions of the lab. A detailed description of ERDE was presented by Losada and Crestani in [29]. ERDE is an error measure that incorporates a penalty for delayed correct alerts (true positives). The penalty increases with the delay in issuing the alert, measured by the number of user posts 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 [32] 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 [33]. 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):

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

[3]Observe that Sadeque et al (see [33], 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.

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

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

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 [33], 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 3 plots how the latency penalty increases with the number of observed writings.

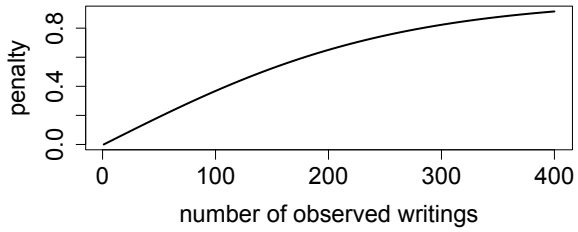

Figure 3: Latency penalty increases with the number of observed writings (k_u)

The system's overall speed factor is computed as:

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

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:

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

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 $ERDE$.

[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.

Table 7
Task 2 (anorexia): 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)
BioNLP-IISERB[34]	5	10	09:39
GVIS	5	352	3 days 12:36
Riewe-Perla [35]	5	2001	2 days 11:25
UNSL [36]	3	2001	07:00
UMU [37]	5	2001	06:34
COS-470-Team-2	5	1	-
ELiRF-UPV [38]	4	2001	12:27
NLP-UNED [39]	5	2001	09:40
SINAI [22]	5	2001	3 days 23:49
APB-UC3M [28]	2	2001	6 days 21:34

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, that is post or comment), 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.
- 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 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 indicates the degree of automation of each team's algorithms. Many of the submitted runs processed the entire thread of messages (2001), but a few variants stopped earlier. Five teams processed the thread of messages reasonably fast (less than a day for processing the entire history of user messages). The rest of the teams took several days to run the whole process.
Table 8 reports the decision-based performance achieved by the participating teams. In terms of $F1$ and latency-weighted $F1$, the best performing team was NLP-UNED [39] (run 1), while Riewe-Perla [35] was the team that submitted the best run (run 0) in terms of the ERDE metrics. 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 anorexia as early as possible.
Table 9 presents the ranking-based results. UNSL [36] (run 1) obtained the best overall values after only one writing, while NLP-UNED [39](run 3) obtained the highest scores after 100 writings. These two teams also contributed the best performing variants for the 500 and 1000 cutoffs.

Table 8
Decision-based evaluation for Task 2

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$lat\text{-}weigh\ F1$
BioNLP-IISERB	0	0.53	0.23	0.32	0.10	0.09	2.00	**1.00**	0.32
BioNLP-IISERB	1	0.54	0.75	0.62	0.08	0.04	4.00	0.99	0.62
BioNLP-IISERB	2	0.58	0.16	0.25	0.10	0.10	**1.00**	**1.00**	0.25
BioNLP-IISERB	3	0.67	0.51	0.58	0.08	0.06	3.00	0.99	0.58
BioNLP-IISERB	4	**0.73**	0.62	0.67	0.08	0.05	4.00	0.99	0.66
GVIS	0	0.12	**1.00**	0.21	0.12	0.10	**1.00**	**1.00**	0.21
GVIS	1	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	2	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	3	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	4	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
Riewe-Perla	0	0.45	0.97	0.62	**0.07**	**0.02**	6.00	0.98	0.60
Riewe-Perla	1	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	2	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	3	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	4	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
UNSL	0	0.35	0.99	0.52	0.14	0.03	12.00	0.96	0.49
UNSL	1	0.42	0.96	0.59	0.14	0.03	12.00	0.96	0.56
UNSL	2	0.42	0.97	0.59	0.14	0.03	12.00	0.96	0.56
UMU	0	0.14	0.99	0.25	0.20	0.09	18.00	0.93	0.23
UMU	1	0.15	0.99	0.26	0.19	0.09	27.00	0.90	0.24
UMU	2	0.14	0.99	0.25	0.20	0.09	19.00	0.93	0.23
UMU	3	0.15	0.99	0.27	0.19	0.09	28.00	0.90	0.24
UMU	4	0.16	0.98	0.27	0.19	0.10	35.50	0.87	0.23
COS-470-Team-2	0	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	1	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	2	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	3	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	4	0.00	0.00	0.00	0.12	0.12			
ELiRF-UPV	0	0.43	0.99	0.60	0.10	0.04	12.00	0.96	0.57
ELiRF-UPV	1	0.41	**1.00**	0.58	0.10	0.04	12.00	0.96	0.56
ELiRF-UPV	2	0.32	0.99	0.49	0.12	0.04	10.00	0.96	0.47
ELiRF-UPV	3	0.43	0.99	0.60	0.11	0.04	15.00	0.94	0.57
NLP-UNED	0	0.64	0.97	0.77	0.09	0.04	13.00	0.95	0.73
NLP-UNED	1	0.67	0.97	**0.79**	0.09	0.04	14.00	0.95	**0.75**
NLP-UNED	2	0.63	0.97	0.76	0.09	0.04	12.00	0.96	0.73
NLP-UNED	3	0.63	0.98	0.77	0.09	0.03	11.00	0.96	0.74
NLP-UNED	4	0.63	0.97	0.76	0.09	0.04	14.00	0.95	0.72
SINAI	0	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	1	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	2	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	3	0.12	**1.00**	0.21	0.13	0.10	2.00	**1.00**	0.21
SINAI	4	0.12	**1.00**	0.21	0.13	0.10	2.00	**1.00**	0.21
APB-UC3M	0	0.17	0.99	0.28	0.15	0.08	9.00	0.97	0.28
APB-UC3M	1	0.15	0.99	0.26	0.13	0.09	2.00	**1.00**	0.26

4. Task 3: Measuring the Severity of Eating Disorders

The objective of the task is to estimate the severity of various symptoms related to the diagnosis of eating disorders. Participants were provided with a thread of user submissions to work with. For each user, a history of posts and comments from Social Media was given, and participants had to estimate the user's responses to a standardized eating disorder questionnaire based on the evidence found in the history of posts/comments.

The questionnaire used in the task is derived from the Eating Disorder Examination Questionnaire (EDE-Q)[6], which is a self-reported questionnaire consisting of 28 items. It is adapted from the semi-structured interview Eating Disorder Examination (EDE)[7] [40]. For this task, we focused on questions 1-12 and 19-28 from the EDE-Q. This questionnaire is designed to assess various aspects and severity of

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

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
BioNLP-IISERB	0	0.10	0.19	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	1	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	2	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	3	0.10	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	4	0.20	0.21	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	0	0.40	0.37	0.40	0.20	0.18	0.23	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	1	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	2	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	3	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	4	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
Riewe-Perla	0	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	1	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	2	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	3	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	4	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
UNSL	0	0.90	0.81	0.63	**1.00**	**1.00**	0.81	**1.00**	**1.00**	0.77	**1.00**	**1.00**	0.76
UNSL	1	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	0.80	0.90	0.81	0.69	0.80	0.88	0.72
UNSL	2	0.40	0.38	0.42	0.90	0.92	0.71	0.80	0.85	0.69	0.80	0.84	0.68
UMU	0	0.20	0.12	0.14	0.10	0.06	0.03	0.00	0.00	0.05	0.20	0.21	0.12
UMU	1	0.20	0.12	0.14	0.10	0.06	0.03	0.00	0.00	0.05	0.20	0.21	0.12
UMU	2	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
UMU	3	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
UMU	4	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
COS-470-Team-2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ELiRF-UPV	0	0.20	0.12	0.14	0.20	0.13	0.14	0.20	0.13	0.14	0.20	0.13	0.14
ELiRF-UPV	1	0.10	0.19	0.17	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.19	0.11
ELiRF-UPV	2	0.10	0.07	0.13	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.06	0.10
ELiRF-UPV	3	0.00	0.00	0.11	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.06	0.10
NLP-UNED	0	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	0.91	**1.00**	**1.00**	0.91
NLP-UNED	1	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	**0.92**	**1.00**	**1.00**	**0.92**
NLP-UNED	2	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	0.91	**1.00**	**1.00**	0.91
NLP-UNED	3	**1.00**	**1.00**	0.45	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.89
NLP-UNED	4	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
SINAI	0	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	1	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	2	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	3	0.00	0.00	0.07	0.10	0.07	0.06	0.00	0.00	0.07	0.00	0.00	0.07
SINAI	4	0.00	0.00	0.07	0.10	0.07	0.06	0.00	0.00	0.07	0.00	0.00	0.07
APB-UC3M	0	0.00	0.00	0.03	0.40	0.56	0.26	0.00	0.00	0.09	0.00	0.00	0.13
APB-UC3M	1	0.10	0.06	0.07	0.00	0.00	0.18	0.00	0.00	0.10	0.00	0.00	0.08

features associated with eating disorders. It includes four subscales: Restraint, Eating Concern, Shape Concern, and Weight Concern, along with a global score. Table 10 shows an excerpt of the EDE-Q.

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..`

Table 10: Eating Disorder Examination Questionarie (continued)

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

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

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

⋮

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)

Table 10: Eating Disorder Examination Questionarie (continued)

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

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

The primary objective of this task was to explore the possibility of automatically estimating the severity of multiple symptoms related to eating disorders. The algorithms are required to estimate the user's response to each individual question based on their writing history. To evaluate the performance of the participating systems, we collected questionnaires completed by Social Media users along with their corresponding writing history. The user-completed questionnaires serve as the ground truth against which the responses provided by the systems are evaluated.
During the training phase, participants were provided with data from 28 users from the 2022 edition and 46 users from the 2023 edition. 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 18 new users for whom the participating systems had to generate results. The results were expected to follow the following specific file structure:

```
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.

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	all 0s	3.790	0.813	4.254	4.472	3.869	4.479	4.363	3.361
baseline	all 6s	1.937	0.551	3.018	3.076	3.352	2.868	3.029	2.472
baseline	average	1.965	0.884	1.973	2.337	2.486	1.559	2.002	1.783
APB-UC3M [28]	0	2.003	0.869	2.142	2.647	2.253	1.884	2.101	1.823
DSGT [27]	0	1.965	0.588	1.713	2.211	2.321	1.969	1.944	2.117
RELAI [23]	0	2.331	0.914	2.243	2.394	2.222	2.324	2.340	**1.812**
RELAI	1	2.346	0.917	2.237	2.507	2.199	2.216	2.328	1.836
RELAI	2	2.758	0.934	2.885	2.883	2.767	3.126	3.061	2.171
RELAI	3	2.356	0.775	2.700	2.928	3.266	2.106	2.821	2.310
RELAI	4	2.851	0.884	2.979	3.159	2.784	3.150	3.068	2.336
SCaLAR-NITK [42]	0	1.912	0.591	**1.643**	2.495	2.713	1.568	**1.536**	2.098
SCaLAR-NITK	1	1.980	0.664	1.972	2.570	2.562	**1.553**	1.960	2.066
SCaLAR-NITK	2	1.879	**0.568**	1.942	2.158	2.477	2.222	2.245	2.364
SCaLAR-NITK	3	1.932	0.586	1.868	**2.117**	2.430	2.046	2.242	2.407
SCaLAR-NITK	4	**1.874**	0.672	1.820	2.292	**2.140**	1.557	1.880	2.061
UMU [37]	0	2.366	0.798	2.833	3.261	3.285	2.659	2.771	2.218
UMU	1	2.227	0.859	2.286	2.326	2.911	2.142	2.560	2.026

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

- **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 reports the results obtained by the participants in this task. In order to provide some 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 the 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 equals 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 4 by SCaLAR-NITK [42]. This team also got the best MZOE (run 2), the best MAE_{macro} (run 0), the best GED (run 3), the best RS (run 4), the best ECS (run 1), and the best SCS (run 0). The best WCS, instead, was achieved by team RELAI [23] (run 0). In some cases the best participating system was not better that some of the baselines (e.g., lowest MZOE is the "all 6s" baseline).

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 2024 working notes proceedings for the participants.

Table 12
eRisk 2024 participants.

team	Task 1 #runs	Task 2 #runs	Task 3 #runs
APB-UC3M [28]	5	2	1
BioNLP-IISERB [34]		5	
COS-470-Team-2		5	
DSGT [27]	5		1
ELiRF-UPV [38]		4	
GVIS	1	5	
MeVer-REBECCA [26]	2		
MindwaveML [25]	3		
NLP-UNED [39]		5	
NUS-IDS [24]	5		
RELAI [23]	5		5
Riewe-Perla [35]		5	
SCaLAR-NITK [42]			5
SINAI [22]	2	5	
ThinkIR	1		
UMU [37]		5	2
UNSL [36]		3	

APB-UC3M [28]. The APB-UC3M team, affiliated with Universidad Carlos III de Madrid (UC3M) in Spain, participated in the three tasks of the eRisk 2024 challenge. For Task 1, which involved searching for symptoms of depression, the team employed sentence similarity models to compare BDI items with paragraphs, in conjunction with a RoBERTa classifier. They also explored ensemble methods combining these approaches. In Task 2, focused on the early detection of anorexia, the team used an ensemble model comprising three classification algorithms. They generated embeddings using BART and Doc2Vec models and utilized these embeddings as input for three traditional classifiers: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). For Task 3, which involved measuring the severity of signs of eating disorders, the team fine-tuned a neural network model. This model included an embedding layer, a fully connected layer, and a ReLU activation function, and it was trained to predict the 22 categories of the Eating Disorder Examination (EDE) interview.

BioNLP-IISERB [34]. The BioNLP-IISERB team, affiliated with the Indian Institute of Science Education and Research, Bhopal, participated in Task 2 of the eRisk 2024 challenge. The team's approach involved a combination of various classification methods and feature engineering techniques to identify signs of anorexia from the provided texts. They utilized both bag-of-words features and transformer-based embedding methods. For classification, they employed Random Forest, Adaptive Boosting, Logistic Regression, Support Vector Machine (SVM), and transformer-based classifiers. Their experimental analysis revealed that the best performance was achieved using SVMs and an Adaboost classifier, particularly with TF-IDF and entropy-based weighting strategies. Some experimental runs achieved F1 scores higher than 0.65, indicating the potential of these frameworks to identify textual patterns indicative of anorexia. Despite the promising results, the complexity of the task suggests there is room for future improvements.

DSGT [27]. The DSGT team, from the Georgia Institute of Technology, participated in Tasks 1 and 3 of the eRisk 2024 challenge. For Task 1, they developed two distinct pipelines to detect signs of depression. Their approach combined traditional NLP techniques, such as TF-IDF, with vector-based models. Specifically, they constructed a logistic regression classifier, treating the 21 symptoms as targets

for a multiclass classification problem. The results on the hidden test set demonstrated that vector models and transformer-based models could achieve notable performance on information retrieval metrics, even without advanced sentence filtering and fine-tuning. For Task 3, the team employed simpler models, including XGBoost and Random Forests, which showed better performance on smaller datasets.

ELiRF-UPV [38]. The ELiRF-VRAIN team, affiliated with the Valencian Research Institute for Artificial Intelligence (VRAIN) at Universitat Politècnica de València, participated in Task 2. Their work involved three distinct approaches: a Support Vector Machine (SVM) and two pre-trained Transformer models. Among the Transformer models, one approach utilized BERT-like models, while the other employed LongFormer models to expand the context when making decisions. To balance the training example set, the authors proposed a data augmentation method, which yielded positive results in augmenting examples during the training process.

MeVer-REBECCA [26]. The REBECCA team, affiliated with the Information Technologies Institute at the Centre for Research and Technology Hellas (CERTH) in Thessaloniki, Greece, participated in Task 1 of the eRisk 2024 challenge, which focused on searching for symptoms of depression. Their approach involved a combination of ranking sentences using cosine similarity and Transformer embeddings, with refinement through a Large Language Model (LLM), specifically ChatGPT-4. The process began with text pre-processing and dataset cleaning, discarding sentences not related to the authors and considering relevant sentences only if they reflected the author's state surrounding a symptom. They conducted keyword matching with sentences indicating self-reference. Following this, the team used sentence ranking with BGE-M3 (Multi-Linguality, Multi-Functionality, and Multi-Granularity) and the questionnaire answers.

MindwaveML [25]. The MindwaveML team, from the University of Bucharest, participated in Task 1. The team leveraged a paraphrasing model to match sentences with BDI texts. Specifically, they encoded the four alternative responses to each of the 21 BDI symptoms into dense embeddings using the paraphrase-MiniLM-L12-v2 model. These embeddings captured the semantic information contained in each response. The sentences from the Reddit corpus were also encoded with paraphrase-MiniLM-L12-v2. The cosine similarity between each sentence and the BDI responses was then computed. Additionally, the team incorporated features to ensure that the sentences contained first-person expressions, ensuring relevance to the individual's state. The resulting set of features was fed to standard learning algorithms.

NLP-UNED [39]. The NLP-UNED team, from UNED in Madrid, participated in Task 2. Their system comprised several steps, starting with an initial embedding representation using sentence encoders. This was followed by a relabelling process based on Approximate Nearest Neighbors (ANN) techniques to generate a training dataset annotated at the message level instead of the user level. The encoding process was further refined with fine-tuning based on contrastive learning, aiming to maximize the distance between embeddings belonging to different classes. For classification, the team also employed ANN techniques, combined with rules and heuristics to expand the number of messages considered from each user when making the final decision. Their system achieved the best results in both the decision-based evaluation and in the ranking-based evaluation.

NUS-IDS [24]. The NUS-IDS team, affiliated with the National University of Singapore's Integrative Sciences and Engineering Programme and the Institute of Data Science, participated in Task 1 of the eRisk 2024 challenge. The team's approach involved ranking candidate sentences for depression symptoms by their average similarity to a predefined set of training sentences. Utilizing methods for computing dense representations of sentences, the team calculated the score of a test sentence as the average cosine similarity between the test sentence and each sentence in a set of training sentences associated with a specific symptom. The authors experimented with different configurations of this algorithm, employing various models for dense representation computation and different sets of training sentences. This approach allowed the NUS-IDS team to effectively rank sentences by their relevance to depression symptoms, leveraging both similarity metrics and the robustness of multiple model configurations.

RELAI [23]. The RELAI team, from Université du Québec à Montréal, Canada, and McMaster University, Canada, participated in Tasks 1 and 3. For Task 1, which involved searching for symptoms of

depression, the team approached it as a multilabel classification task. They utilized feed-forward neural networks with contextual embeddings to mine sentences relevant to each item in a standard depression questionnaire from a large set of social media sentences. Their methods aimed to be lightweight, minimizing computational costs and infrastructure needs. In Task 3, the team used BERTopic to extract the 16 most correlated topics with signs of eating disorders as features for prediction. They employed feed-forward neural networks with topic probabilities as inputs to automatically fill out a standard eating disorder questionnaire based on social media writing histories. The authors noted significant room for improvement, particularly in exploring different representations of writing history and improving model calibration for classification transformations.

Riewe-Perla (MHRec) [35]. The Poznan University of Economics and Business team participated in Task 2. Their approach involved merging language models with recommender systems to analyze and predict if recommended content originated from individuals with mental health conditions. The team's model was built on document embeddings, user embeddings, and a recommendation engine using the sentence transformer architecture (SBERT). They employed a hybrid recommendation method (LightFM) that leveraged both document and user embeddings to flag publications indicating mental health challenges. The system aimed to facilitate fast classification of new messages, determining as early as possible whether an individual was suffering from anorexia.

SCaLAR-NITK [42]. Team SCaLAR-NITK, from the National Institute of Technology Karnataka, Surathkal, participated in Task 3. The team employed a range of standard techniques across 21 different models—one for each symptom question. Their first approach utilized Support Vector Machine (SVM) classifiers with input word embeddings constructed using the traditional TF-IDF method. In the second approach, they again used SVMs but leveraged pre-trained Word2Vec embeddings to model both users and questions, aggregating the question embeddings with each user publication. To address response imbalance, they employed back-translation. Their final method followed the second approach but incorporated Principal Component Analysis (PCA) for dimensionality reduction of embeddings. Their methods performed well, achieving the best results in 7 out of the 8 evaluated metrics.

SINAI [22]. The SINAI team, a collaborative effort between the Computer Science Department of Universidad of Jaén (Spain) and Instituto Nacional de Astrofísica, Óptica y Electrónica (Mexico), participated in Tasks 1 and 2. For Task 1, one of SINAI's approaches involved training a DistilRoBERTa base model on labeled sentences, with additional data augmentation using the BDI-Sen dataset. Another approach for Task 1 utilized GPT-3 prompts to infer connections between PHQ-8 symptoms and BDI symptoms. For Task 2, the team implemented two transformer-based models trained with causal language modeling, one trained on positive user data and the other on negative user data. This dual-model solution was used to produce perplexity estimates.

UMU Team [37]. The UMU Team, from the University of Murcia (Spain), participated in Tasks 2 and 3. For Task 2, the team proposed a method that classifies user posts by combining the last-layer hidden representation of a BERT-based model with sentiment features extracted from the text. They utilized BERT and RoBERTa models for text representation, along with the Cardiff NLP TweetEval model for sentiment analysis. This approach aimed to capture both the semantic and emotional aspects of the users' posts to detect signs of anorexia. For Task 3, they adopted a fine-tuning approach using a sentence transformer model to compute the similarity between the text of the user and the responses of the EDE-Q questionnaire. This method involved measuring the textual closeness between user posts and the EDE-Q questions to assess the severity of eating disorder symptoms.

UNSL [36]. The UNSL team, from Universidad Nacional de San Luis (Argentina), participated in Task 2 with a solution named CPI-DMC, focusing on precision and speed independently, as well as a time-aware approach where both objectives are tackled together. The first approach aimed to balance identifying positive users and minimizing the decision-making time, consisting of two separate components: a Classifier with Partial Information (CPI) and another for Deciding the Moment of the Classification (DMC). The second approach aimed to optimize both objectives simultaneously by incorporating time into the learning process and using ERDE as the training objective. To implement this, they included a [TIME] token in the representations, integrating temporal metrics to validate and select the optimal models. Their methods achieved good results for the ERDE50 metric and ranking-based metrics, and

demonstrating consistency in solving early risk detection problems.

6. Conclusions

This paper provided an extended overview of eRisk 2024, the eighth edition of the lab, which focused on three types of tasks: symptoms search (Task 1 on depression), early detection (Task 2 on anorexia), and severity estimations (Task 3 on eating disorders). Participants in Task 1 were given a collection of sentences and had to rank them according to their relevance to each of the BDI-II depression symptoms. Participants in Task 2 had sequential access to social media posts and had to send alerts about individuals showing risks of anorexia. In Task 3, participants were given the full user history and had to automatically estimate the user's responses to a standard depression questionnaire.
A total of 87 runs were submitted by 17 teams for the proposed tasks. The experimental results demonstrate the value of extracting evidence from social media, indicating that automatic or semi-automatic screening tools to detect at-risk individuals could be promising. These findings highlight the need 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 (GPC ED431B 2022/33) and the European Regional Development Fund and project PID2022-137061OB-C21 (MCIN/AEI/ 10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by "ERDF A way of making Europe", by the "European Union"). The CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, CITIC is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). 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. David E. Losada also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego) and project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

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, p. 233–256.
[18] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2023: Early risk

prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18–21, 2023, 2023.
[19] 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.
[20] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, 2024.
[21] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An Inventory for Measuring Depression, JAMA Psychiatry 4 (1961) 561–571.
[22] A. M. Mármol-Romero, P. A.-O. Adrián Moreno-Muñoz, K. M. Valencia-Segura, E. Martínez-Cámara, M. García-Vega, A. Montejo-Ráez, SINAI at eRisk@ CLEF 2024: Approaching the Search for Symptoms of Depression and Early Detection of Anorexia Signs using Natural Language Processing, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[23] D. Maupomé, Y. Ferstler, S. Mosser, M.-J. Meurs, Automatically finding evidence, predicting answers in mental health self-report questionnaires , in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[24] B. H. Ang, S. D. Gollapalli, S.-K. Ng, NUS-IDS@eRisk2024: Ranking Sentences for Depression Symptoms using Early Maladaptive Schemas and Ensembles , in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[25] R.-M. Hanciu, MindwaveML at eRisk 2024: Identifying Depression Symptoms in Reddit Users, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[26] A. Barachanou, F. Tsalakanidou, S. Papadopoulos, REBECCA at eRisk 2024: Search for symptoms of depression using sentence embeddings and prompt-based filtering , in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[27] D. Guecha, A. Potdar, A. Miyaguchi, DS@GT eRisk 2024 Working Notes , in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[28] A. P. Bascuñana, I. S. Bedmar, APB-UC3M at eRisk 2024: Natural Language Processing and Deep Learning for the Early Detection of Mental Disorders, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[29] 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.
[30] 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.
[31] 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.
[32] 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).
[33] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: WSDM, ACM, 2018, pp. 495–503.
[34] P. Sarangi, S. Kumar, S. Agrawal, T. Basu, A natural language processing based framework for early detection of anorexia via sequential text processing, in: Working Notes of CLEF 2024 -

Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[35] O. Riewe-Perła, A. Filipowska, Combining Recommender Systems and Language Models in Early Detection of Signs of Anorexia, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[36] H. Thompson, M. Errecalde, A Time-Aware Approach to Early Detection of Anorexia: UNSL at eRisk 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[37] R. Pan, J. A. G. Díaz, T. B. Beltrán, R. Valencia-Garcia, UMUTeam at eRisk@CLEF 2024: Fine-Tuning Transformer Models with Sentiment Features for Early Detection and Severity Measurement of Eating Disorders , in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[38] A. C. Segarra, V. A. Esteve, A. M. Marco, L.-F. H. Oliver, ELiRF-VRAIN at eRisk 2024: Using Long-Formers for Early Detection of Signs of Anorexia, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[39] H. Fabregat, D. Deniz, A. Duque, L. Araujo, J. Martinez-Romo, NLP-UNED at eRisk 2024: Approximate Nearest Neighbors with Encoding Refinement for Early Detecting Signs of Anorexia, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.
[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] S. Prasanna, A. S. Gulati, S. Karmakar, M. Y. Hiranmayi, A. K. Madasamy, Measuring the severity of the signs of eating disorders using machine learning techniques, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, September 9-12, 2024.