eRisk 2024: Depression, Anorexia, and Eating Disorder Challenges

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, A Coruña, Spain
{javierparapar,patricia.martin.rodilla}@udc.es
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago, Spain
david.losada@usc.es
[3] Faculty of Informatics, Universitá della Svizzera italiana (USI), Lugano, Switzerland
fabio.crestani@usi.ch

Abstract. In 2017, we launched eRisk as a CLEF Lab to encourage research on early risk detection on the Internet. Since then, thanks to the participants' work, we have developed detection models and datasets for depression, anorexia, pathological gambling and self-harm. In 2024, it will be the eighth edition of the lab, where we will present a revision of the sentence ranking for depression symptoms, the third edition of tasks on early alert of anorexia and eating disorder severity estimation. This paper outlines the work that we have done to date, discusses key lessons learned in previous editions, and presents our plans for eRisk 2024.

1 Introduction

The eRisk Lab[1] is an ongoing project focused on evaluating early risk detection on the internet, with a particular emphasis on health and safety issues. Since its pilot edition in 2017 in Dublin [6], it has been a part of CLEF. Throughout its various editions [6–9,11–13], many collections and models have been presented under the eRisk umbrella, and the presented dataset construction approaches and evaluation strategies can be applied to different types of risks.

Our interdisciplinary Lab addresses tasks that require and combine from information retrieval, computational linguistics, machine learning, and psychology knowledge. Diverse experts have collaborated to design monitoring models for critical societal problems. These models could be used, for example, to alert when someone exhibits signs of suicidal thoughts on social media. Previous eRisk editions have addressed issues like depression, eating disorders, gambling, and self-harm detection.

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

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14612, pp. 474–481, 2024.
https://doi.org/10.1007/978-3-031-56069-9_65

eRisk has introduced early alert, sentence ranking for risk symptoms, and severity estimation tasks. Early risk tasks (Sect. 2.1) involve predicting risks by analyzing a temporal text stream (e.g., social media posts) and accumulating evidence to make decisions about specific risks, such as the development of depression. In the severity estimation challenges (Sect. 2.2), participants use all user writings to compute a detailed estimate of symptoms of a specific risk, filling out a standard questionnaire as real users would do. Last year, we presented the sentence ranking for signs of depression (Sect. 2.3). This new type of task complements the other two by challenging participants to rank sentences from a collection of user writings according to their relevance to the symptoms of a specific risk.

2 A Brief History of eRisk

In the inaugural eRisk edition in 2017 [6], the sole pilot task focused on early depression risk detection. Data was released in weekly chunks, and participants had to submit predictions after each release. The demanding nature of this process led to only eight groups out of thirty completing the tasks by the deadline. Evaluation methods and metrics were based on those defined in [5].

In 2018 [7], the same setup was continued, featuring the early detection of depression and introducing a new task for early detection of anorexia. Task 1 on depression received 45 system submissions, while for anorexia we received 35.

In 2019 [8], a significant change occurred as the release of user posts became more fine-grained using a server. Two early risk detection tasks continued (anorexia and self-harm), and a new task on severity estimation for depression was introduced, involving clinically validated questionnaires. The number of submissions for these tasks was 54, 33, and 33 for tasks 1, 2, and 3, respectively. In 2020 [9], the early detection of self-harm and the estimation of depression symptom severity tasks persisted. There were 46 system submissions for the early risk task and 17 for the severity estimation task.

The year 2021 [11] saw the introduction of three tasks, with the third edition of the early self-harm detection and depression symptom severity tasks. Additionally, a new task focused on early detection in the domain of pathological gambling was introduced, receiving 115 runs from 18 teams out of 75 registered. In 2022 [12], the early risk detection of pathological gambling task continued, along with the first edition of early depression risk detection under the new fine-grained setup. Another severity estimation task was presented, focusing on eating disorders and using a standard questionnaire. In total, the proposed tasks received 117 runs from 18 teams.

In 2023 [13], a new task was introduced, focusing on locating markers in sentences for 21 depression symptoms as defined by the BDI-II questionnaire. This marked the first edition of the sentence ranking task. The three-year cycle for early alert of pathological gambling was closed and we also ran the second edition of the eating disorder severity estimation task under the EDE-Q questionnaire, with 105 runs submitted by 20 teams for the proposed tasks.

Over these seven years, eRisk has received a stable number of active participants, slowly placing the Lab as a reference forum for early risk research. We summarised the eRisk experience and the best models presented so far in our recent book [4].

2.1 Early Risk Prediction Tasks

The initial challenges revolved around early risk prediction in various domains (such as depression, anorexia, and self-harm). In each edition, the teams analyzed social media writings (posts or comments) sequentially in chronological order to detect signs of risk as early as possible. All shared tasks in different editions sourced their data from the social media platform Reddit.

Reddit users tend to write prolifically, often with posts spanning several years. Many communities (subreddits) on Reddit focus on mental health disorders, providing a valuable resource for obtaining the writing history of redditors to build eRisk datasets [5]. In these datasets, redditors are categorized into a positive class (e.g., individuals with depression) and a negative class (control group). To identify positive redditors, the methodology developed by Coppersmith and colleagues [3] was followed. For instance, in the case of identifying positive redditors for depression, the writings were searched for explicit strings (e.g., "Today, I was diagnosed with depression") indicating a diagnosis. Notably, phrases like "I am anorexic" or "I have anorexia" were not considered explicit affirmations of a diagnosis. This semi-supervised method has been used to extract information about patients diagnosed with different conditions since 2020, with the assistance of the Beaver tool [10] for labeling positive and negative instances.

Regarding the evaluation methodology, the first edition of eRisk introduced a new metric called ERDE (Early Risk Detection Error) to measure early detection [5]. ERDE differs from standard classification metrics as it takes into account prediction latency, considering both the correctness of the binary decision and the latency. The original ERDE metric quantified latency by counting the number of posts (k) processed before reaching a decision. In 2019, an alternative metric for early risk prediction, $F_{latency}$, was adopted, as proposed by Sadeque et al. [14]. Starting in 2019, with the release of user texts at the writing level, user rankings were generated based on participants' estimated degree of risk. These rankings have been evaluated using common information retrieval metrics, including P@10 and nDCG [8].

2.2 Severity Level Estimation Tasks

In 2019, we introduced a new task focused on estimating the severity of depression, a task that continued in 2020 and 2021. In 2022 we ran the severity estimation on the eating disorder domain, where participants were required to automatically complete the EDE-Q questionnaire. In these tasks, participants had access to the writing history of some redditors who volunteered to fill out the standard questionnaires. Their challenge was to develop models that could answer each

of the questionnaire's questions based on the evidence found in the provided writings.

For depression assessment, we used Beck's Depression Inventory (BDI-II) [1], which consists of 21 questions related to the severity of depression signs and symptoms, each with four alternative responses corresponding to different severity levels (e.g., loss of energy, sadness, and sleeping problems). For eating disorders, we used questions 1–12 and 19–28 from the Eating Disorder Examination Questionnaire (EDE-Q) [2].

To create the ground truth, we compiled surveys completed by social media users, along with their writing history. Given the unique nature of the task, we introduced new evaluation measures to assess the participants' estimations. In the case of depression, four metrics were defined: Average Closeness Rate (ACR), Average Hit Rate (AHR), Average DODL (ADODL), and Depression Category Hit Rate (DCHR). Detailed descriptions of these metrics can be found in [8]. In the case of eating disorders, we adopted new metrics in the previous year, including Mean Zero-One Error (MZOE), Mean Absolute Error (MAE), Macroaveraged Mean Absolute Error (MAE_{macro}), Global ED (GED), and the corresponding Root Mean Square Error (RMSE) for four sub-scales: Restraint, Eating Concern, Shape Concern, and Weight Concern [12].

2.3 Sentence Ranking for Symptoms of Risk Tasks

In the 2023 edition, Task 1 presented a novel challenge, focusing on the creation of sentence rankings based on their relevance to specific symptoms of depression. Participants were instructed to rank sentences extracted from user writings based on their relevance to the 21 standardized symptoms as outlined in the BDI-II Questionnaire [1]. In this context, a sentence was considered relevant to a particular symptom if it provided information about the user's condition related to that symptom. It's crucial to highlight that a sentence could be considered relevant even if it conveyed positive information about the symptom. For example, a sentence like "I feel quite happy lately" should still be regarded as relevant for symptom 1, which is "Sadness" in the BDI-II. Using participants' results and *top-k* pooling as document adjudication model, we created the relevance judgements with expert assessors. The ranking-based evaluation was conducted using Mean Average Precision (MAP), mean R-Precision, mean Precision at 10, and mean nDCG at 1000.

2.4 Results

According to the CLEF tradition, Labs' Overview and Extended Overview papers compile the summaries and critical analysis of the participants' systems results [6–9,11–13].

Over the course of eleven editions of early detection tasks for four mental health disorders, we have seen a diverse array of models and methods, with most participants primarily focusing on improving classification accuracy on training data, rather than considering the accuracy-delay trade-off that's crucial

for timely detection. Notably, we've observed varying system performance across different disorders over the years. For instance, anorexia and pathological gambling tasks appear to be more manageable than depression detection, possibly due to differences in available training data and the nature of the disorders.

Our observations suggest that the likelihood of patients leaving traces of their condition in their social media language may vary depending on the illness. Nonetheless, the results demonstrate a consistent pattern of participants improving detection accuracy from edition to edition, which encourages us to continue supporting research in text-based early risk detection on social media.

Furthermore, some participants have shown promising results in developing automatic or semi-automatic screening systems for predicting the severity of specific risks. The results also suggest that analyzing the entire user's writing history can be a complementary technique for extracting indicators or symptoms related to the disorder, particularly for depression, where some systems achieved a 40% hit rate in answering the BDI questions in the same way as real users. In the case of the eating disorder questionnaire, results in the second edition, with participants using training data, showed improvement over the previous edition.

In the inaugural sentence ranking task, performance varied among the 37 runs, but the best team achieved promising results (best nDCG: 0.596, best P@10: 0.861).

3 The Tasks of eRisk 2024

The outcomes of previous editions have encouraged us to continue the Lab in 2024 and examine the interaction between text-based screening from social media and risk prediction and estimation. The following is the task breakdown for our plans for this year:

3.1 Task 1: Search for Symptoms of Depression

As in the 2021 task, this challenge will involve ranking sentences from a collection of user writings based on their relevance to each of the 21 symptoms of depression outlined in the BDI-II questionnaire. A sentence will be considered relevant to a BDI symptom if it provides information about the user's condition related to that symptom. We will provide a dataset of tagged sentences along with the BDI-II questionnaire. Participants are free to choose their strategy for generating queries based on the BDI symptom descriptions in the questionnaire. Each system will submit 21 sentence rankings, one for each BDI item.

Once we receive submissions from the participating teams, we will create relevance judgments with the assistance of three human assessors using top-k pooling. These resulting *qrels* will be used to evaluate the systems using standard ranking metrics like MAP, nDCG, among others. This newly annotated corpus of sentences will be a valuable resource with numerous applications beyond eRisk.

In this second edition of the task, we will provide participants with the data from the previous year. Additionally, building on the experience gained in 2023,

we plan to include some context to assess the relevance of the target sentences. Specifically, relevance will be determined based on the target sentence along with the preceding and following sentences from the original corpus.

3.2 Task 2: Early Detection of Anorexia

The challenge focuses on the sequential processing of evidence to detect early signs of anorexia as quickly as possible. Texts are processed in the order of their creation, allowing systems that excel at this task to be applied for sequential monitoring of user interactions in blogs, social networks, or other online media.

The task is divided into two stages. In the training stage, participating teams will have access to a training server, where we will release the complete history of writings for a set of training users. The training data will be derived from the 2018 and 2019 editions. The test stage involves a period during which participants must connect to our server[2] and iteratively retrieve user writings and submit their responses.

3.3 Task 3: Measuring the Severity of the Signs of Eating Disorders

The task involves assessing an individual's level of eating disorder based on their historical written submissions. Participants are required to use automated solutions to complete a standard eating disorder questionnaire based on the complete history of user's writings.

The Eating Disorder Examination Questionnaire (EDE-Q) is used to evaluate the range and severity of features associated with eating disorders. It consists of 28 items divided into four subscales: restraint, eating concern, shape concern, and weight concern, in addition to a global score [2]. Using the user's written history, algorithms must estimate the user's responses to each individual item.

We will gather questionnaires filled out by social media users, along with their written histories (collected immediately after the user completes the questionnaire). These user-filled questionnaires serve as the ground truth and will be used to evaluate the quality of responses provided by participating systems. Participants will have access to training data from 2022 and 2023.

4 Conclusions

The results achieved in eRisk and the engagement of the research community inspire us to keep introducing new challenges related to risk identification in Social Media. We extend our heartfelt gratitude to all participants for their contributions to the success of eRisk. We strongly encourage research teams to continue refining and developing new models for upcoming tasks and risks. Despite the time and effort required to create these resources, we firmly believe that the societal benefits far outweigh the associated costs.

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

Acknowledgements. The first and second authors thank the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G/01, 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 and the project PID2022-137061OB-C21 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; suppported by the European Regional Development Fund). The third author thanks the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (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. The third author thanks the financial support obtained from: i) project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; suppported by the European Regional Development Fund) and ii) project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego). The first, second, and third author also thank the funding of 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).

References

1. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. JAMA Psychiatry **4**(6), 561–571 (1961)
2. Carey, M., Kupeli, N., Knight, R., Troop, N.A., Jenkinson, P.M., Preston, C.: Eating disorder examination questionnaire (EDE-Q): norms and psychometric properties in UK females and males. Psychol. Assess. **31**(7), 839 (2019)
3. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: ACL Workshop on Computational Linguistics and Clinical Psychology (2014)
4. Crestani, F., Losada, D.E., Parapar, J. (eds.): Early Detection of Mental Health Disorders by Social Media Monitoring. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04431-1
5. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016, Evora, Portugal (2016)
6. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2017: clef lab on early risk prediction on the internet: experimental foundations. In: Jones, G.J., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 346–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_30
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: early risk prediction on the internet. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 343–361. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_30
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 early risk prediction on the internet. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 340–357. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_27

9. Losada, D.E., Crestani, F., Parapar, J.: Overview of of eRisk 2020: early risk prediction on the internet. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 272–287. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_20
10. Otero, D., Parapar, J., Barreiro, Á.: Beaver: efficiently building test collections for novel tasks. In: Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, 6–9 July 2020 (2020). https://ceur-ws.org/Vol-2621/CIRCLE20_23.pdf
11. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of eRisk 2021: early risk prediction on the internet. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 324–344. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_22
12. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of eRisk 2022: early risk prediction on the internet. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 233–256. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_18
13. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of eRisk 2023: early risk prediction on the internet. In: Arampatzis, A., et al. (eds.) CLEF 2022. LNCS, vol. 14163, pp. 294–315. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_22
14. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, pp. 495–503. ACM, New York (2018)