eRisk 2023: Depression, Pathological Gambling, 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 2023, it will be the seventh edition of the lab, where we will present a new type of task on sentence ranking for depression symptoms. This paper outlines the work that we have done to date, discusses key lessons learned in previous editions, and presents our plans for eRisk 2023.

1 Introduction

The eRisk Lab[1] is a forum for exploring evaluation methodologies and effectiveness metrics related to early risk detection on the Internet (with past challenges particularly focused on health and safety). Since the pilot edition in 2017 in Dublin, we have been part of CLEF, the Conference and Labs of the Evaluation Forum. Along the different editions [7–10,12,13] many collections and models have been presented under the eRisk banner. Our dataset construction approach and evaluation strategies are broad, meaning they might be applied to various application domains.

Our interdisciplinary Lab addresses tasks touching areas such as information retrieval, computational linguistics, machine learning, and psychology. Participants with heterogeneous expertise collaborate to design monitoring solutions for critical social worrying problems. Ideally, the developed models may be used in systems that, for instance, will alert when someone shows suicidal ideas or on social media. Previous eRisk editions included joint work on depression, eating disorders, gambling, and self-harm detection.

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

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 585–592, 2023.
https://doi.org/10.1007/978-3-031-28241-6_67

So far, eRisk eRisk has proposed early alert and severity estimation tasks. Early risk tasks (Sect. 2.1) involve evidence-building risk prediction. For that, participant systems must automatically analyse a temporal text stream from a source (e.g., social media posts) while accumulating evidence to decide about a specific risk (e.g. developing depression). On the other hand, in the severity estimate challenges (Sect. 2.2), participants use all user writings for computing a fine-grained estimate of the symptoms of a specific risk. With that, models must fill out a standard questionnaire as real users would.

2 A Brief History of eRisk

Since the Lab's inception, we have created numerous reference collections in the field of risk prediction in depression, anorexia, self-harm, and pathological gambling disorders.

In the first eRisk [7], early risk of depression was the only pilot task. This edition released temporal data chunks in sequential order (one chunk per week). Following each release, participants submitted their predictions. This demanding procedure resulted in only eight (up to 5 systems each) of the thirty participating groups completing the tasks by the deadline. The evaluation methodology and metrics were those defined in [6]. In 2018 [8], we maintained the same setup for a continuation of the early detection of depression task and a new one on the early detection of signs of anorexia. Participants submitted 45 systems for Task 1 (depression) and 35 for Task 2 (anorexia).

It was in 2019 [9] when we moved from the weekly chunk release of users' writings to a fine-grained release of user posts using a server. We used that approach for task 1 on early risk detection of anorexia and task 2, a new task on self-harm. Another important change was the introduction of a new task on severity estimation using clinically validated questionnaires. We presented this new kind of task for the case of depression. In this new challenge, the participants received the whole writing history of the users. We received 54, 33, and 33 for tasks 1,2 and 3. In 2020 [10], we continued the tasks of early detection of self-harm and estimating the severity of depression symptoms. Participants submitted 46 system variants for the early risk task and 17 different runs for the severity estimation one.

We proposed three tasks in 2021 [12]. Following our three-year-per-task cycle, we closed the tasks on the early detection of signs of self-harm challenge and the estimation of the severity of the symptoms of depression. We also presented a new domain for early detection, in this case, pathological gambling. We received 115 runs from 18 teams out of 75 registered. In 2022, we continued the task of early risk detection of pathological gambling. We closed the cycle for early risk detection of depression (the first edition under the new fine-grained setup). Additionally, we presented a new severity estimation task using, in this case, a standard questionnaire on eating disorders. The proposed tasks received 117 runs from 18 teams in total.

Over these five years, eRisk has received a steady 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 primary goal of eRisk was to develop practical algorithms and models for tracking social network activity. Most of the presented challenges were early predicting risk in various domains (depression, anorexia, self-harm). They were all organised the same way: the teams had to analyse social media writings (posts or comments) sequentially (in chronological order) to spot signs of risk as early and feasible as possible.

All shared tasks in the different editions were sourced from the social media platform Reddit. It is critical to highlight that data extraction for research purposes is permitted under Reddit's terms of service. Reddit does not permit unauthorised commercial use or redistribution of its content except as authorised by the concept of fair use. eRisk's research activities are an example of fair usage.

Redditors tend to be prolific in writing, being common to have many posts published over several years. There are many communities (subreddits) dedicated to mental health disorders such as depression, anorexia, self-harm, or pathological gambling. We leverage that for obtaining the writing history of redditors (posts or comments) for building the eRisk datasets [6]. All our datasets, redditors are divided into positive class (e.g., depressed) and negative class (control group). To obtain them, we followed Coppersmith and colleagues [3] methodology. For instance, when looking for positive redditors for depression, we searched for in the writings for explicit strings (e.g. "Today, I was diagnosed with depression") about the redditors being diagnosed with depression. For example, "*I am anorexic*", "*I have anorexia*", or "*I believe I have anorexia*" were not deemed explicit affirmations of a diagnosis. We followed this semi-supervised method for extracting information about patients diagnosed with different conditions. Since 2020, we have used Beaver [11], a new tool for labelling positive and negative instances, for aiding us in this task.

In terms of evaluation methodology, in the first edition of eRisk, we presented a new measure called ERDE (Early Risk Detection Error) for measuring early detection [6]. Contrary to standard classification metrics that ignore prediction latency, ERDE considers both the correctness of the (binary) decision and the latency. In the original ERDE, the latency corresponds with the counting of posts (k) processed before reaching the decision. In 2019, we also adopted $F_{latency}$, an alternative assessment metric for early risk prediction proposed by Sadeque et al. [14].

With the introduction of the writing-level release of user texts in 2019, we could produce user rankings by the participants-provided estimated degree of

risk. Since then, we have evaluated those ranks using common information retrieval metrics (for example, P@10 or nDCG) [9].

2.2 Severity Level Estimation Task

In 2019 we introduced a new task on estimating the severity level of depression that we continued in 2020 and 2021. In the last edition, we introduced a new task in the eating disorder domain where participants had to fill out the EDE-Q questionnaire automatically. Those tasks investigate the feasibility and potential methods for automatically measuring the occurrence and severity of various well-known symptoms for the mentioned disorders. In this task, participants had access to the history of writings of some redditors who have volunteered to fill out the standard questionnaire. Participants had to produce models that answered each of the questions of the corresponding standard based on the evidence found in the provided writings.

In the case of depression, we used the Beck's Depression Inventory (BDI) [1]. It presents 21 questions regarding the severity of depression signs and symptoms (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 produce the ground truth, we compiled a series of surveys by social media users with their writing history. Because of the unique nature of the task, we presented new evaluation measures for evaluating the participants' estimations. We defined four metrics in the depression scenario: Average Closeness Rate (ACR), Average Hit Rate (AHR), Average DODL (ADODL), and Depression Category Hit Rate (DCHR), details can be found in [9]. Last year, for the eating disorder results, we also adopted new metrics: 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 the four sub-scales: Restraint, Eating Concern, Shape Concern, Weight Concern [13].

2.3 Results

According to the CLEF tradition, Labs' Overview and Extended Overview papers compile the summaries and critical analysis of the participants' systems results [7–10,12].

So far, ten editions of early detection tasks on four mental health disorders have been celebrated. We have received a diverse range of models and methods. Many of them rely on traditional classification techniques. That is, most participants focused on improving classification accuracy on training data. As this task tries to promote fast-responding models to signs of the disorder, we missed more systems concerned with the accuracy-delay trade-off in general. In any case, we have observed a non-homogeneous system performance along the different disorders over the years. For instance, anorexia and pathological gambling seem to

be more manageable tasks than depression detection. These discrepancies could be attributed to the amount and quality of released training data and the illness itself. We hypothesise that, depending on the illness, patients are more or less likely to leave traces of their social media language. The results reveal a pattern in how participants improved detection accuracy from edition to edition. This pattern encourages us to continue funding research on text-based early risk detection in social media. Furthermore, based on the performance of some participants, automatic or semi-automatic screening systems that predict the onset of specific hazards appears to be within reach.

The results also demonstrate that automatic analysis of the whole user's writing history could be a complementary technique for extracting some indicators or symptoms connected with the disorder when determining disease severity. In the case of depression, for example, where participants had access to training data, some systems had a 40% hit rate (the systems answered 40% of the BDI questions with the exact same response as the real user). Although there is still much room for improvement, this demonstrates that the participants were able to extract some signals from the jumbled social media data. For the eating disorder questionnaire, the results for the first edition are still very modest, considering that participants of the first edition had no access to training data.

The difficulties in locating and adapting measures for these novel challenges have prompted us to develop new metrics for eRisk. Some eRisk participants [14,15], were also engaged in proposing novel modes of evaluation, which is yet another beneficial outcome of the Lab. As commented, we incorporated error metrics in the new severity estimation task last year. Both MAE and RMSE are two widely used metrics in rating prediction for users in recommendation systems [5].

3 The Tasks of eRisk 2023

The outcomes of previous editions have encouraged us to continue the Lab in 2023 and examine the interaction between text-based screening from social media and risk prediction and estimation. The following is the task breakdown for eRisk 2023:

3.1 Task 1: Search for Symptoms of Depression

This is a new type of challenge. The task consists of ranking sentences from a collection of user writings according to their relevance to a depression symptom. The participants will have to provide rankings for the 21 symptoms of depression from the BDI questionnaire [1]. A sentence will be deemed relevant to a BDI symptom when it conveys information about the user's state concerning the symptom. That is, it may be relevant even when it indicates that the user is ok with the symptom.

We would release a sentence-tagged dataset (based on eRisk past data) together with the BDI questionnaire. Participants would be free to decide on

the best strategy to derive queries from describing the BDI symptoms in the questionnaire. After receiving the runs from the participating teams, we would create the relevance judgements with the help of human assessors using pooling. We will use the resulting *qrels* to evaluate the systems with classical ranking metrics (e.g. MAP, nDCG, etc.). This new corpus with annotated sentences would be a valuable resource with multiple applications beyond eRisk.

3.2 Task 2: Early Detection of Pathological Gambling

In 2023 it will be the third edition of the task. It follows the early detection challenge. It consists of sequentially processing pieces of evidence and detecting early traces of pathological gambling as soon as possible. Participants must process Social Media texts in the order the users wrote them. In this way, systems that effectively perform this task could be applied to sequentially monitor user interactions in blogs, social networks, or other types of online media. We will provide the data from 2021 and 2022 as training data. The test stage will consist of a period where the participants have to connect to our server[2]. and iteratively get user writings and send decisions.

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

The task consists of estimating the severity level of the eating disorder given a user history or written submissions. For that, we provide participants with the postings history, and the participants will have to fill out a standard eating disorder questionnaire (based on the evidence found in texts). The EDE-Q assesses the range and severity of features associated with the diagnosis of eating disorders. It is a 28-item questionnaire with four subscales (restrain, eating, concern, shape concern, and weight concern) and a global score [2]. The questionnaires filled by the users (ground truth) will be used to assess the quality of the responses provided by the participating systems. Participants will have training data from last year.

4 Conclusions

The results obtained so far under eRisk and the research community's participation drive us to continue proposing new challenges related to risk identification in Social Media. We sincerely thank all participants for their contributions to eRisk's success. We want to encourage the research teams to keep improving and developing new models for future tasks and dangers. Even while creating new resources is time-consuming, we believe that the societal benefits outweigh the costs.

[2] http://early.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. 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 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 Psychiat. **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, Heidelberg (2022). https://doi.org/10.1007/978-3-031-04431-1
5. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
6. 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)
7. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: Jones, G.J.F., 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
8. 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
9. 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
10. Losada, D.E., Crestani, F., Parapar, J.: Overview 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

11. 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
12. 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
13. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: 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, 5–8 September 2022, Proceedings, pp. 233–256. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-13643-6_18
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)
15. Trotzek, M., Koitka, S., Friedrich, C.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. **32**, 588–601 (2018)