eRisk 2022: Pathological Gambling, Depression, 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 de Compostela, 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. The eRisk 2021 was the fifth edition of the Lab. Since then, we have created a large number of collections for early detection addressing different problems (e.g., depression, anorexia or self-harm). This paper outlines the work that we have done to date (2017, 2018, 2019, 2020, and 2021), discusses key lessons learned in previous editions, and presents our plans for eRisk 2022, which introduces a new challenge to assess the severity of eating disorders.

1 Introduction

As a part of CLEF (Conference and Labs of the Evaluation Forum), 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). Over the previous editions [6–9,11], we have presented a number of testbeds and tools under the eRisk's umbrella. Our dataset construction methodology and evaluation strategies are general and, thus, potentially applicable to different application domains.

Our Lab brings together researchers from various fields (such as information retrieval, computational linguistics, machine learning, and psychology) to interdisciplinary address the presented tasks. Furthermore, participants develop models for solving the eRisk defined challenges that may play a critical role in helping in solving socially worrying problems. For example, when an individual begins broadcasting suicidal thoughts or threats of self-harm on social networks,

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

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 436–442, 2022.
https://doi.org/10.1007/978-3-030-99739-7_54

systems may send warning alerts. Previous eRisk editions proposed shared tasks centred on specific health and security issues, such as depression, anorexia, or self-harm detection.

eRisk proposes two types of challenges: early alert and severity estimation tasks. On early risk tasks, risk prediction is viewed as a sequential process of evidence accumulation. Participant systems must automatically analyse the continuous data flow in a given source (e.g., social media entries). Those algorithms must estimate when and if there is enough aggregated evidence about a specific type of risk during this process. On each shared task, participants have access to temporally organised writing histories and must balance making early alerts (e.g., based on a few social media entries) versus making not-so-early (late) alerts (e.g., evaluating a wider range of entries and only emitting alerts after analysing a larger number of pieces of evidence). On the other hand, severity estimation tasks are concerned with computing a fine-grained estimate of the symptoms of a specific risk based on the entire set of user writings. Participants are challenged to create models that fill out a standard questionnaire the same way that a real user would.

2 Five Years of eRisk

eRisk, a CLEF lab for research on early risk prediction on the Internet, began in 2017 as a forum to lay the experimental groundwork for early risk detection. Our fifth anniversary was last year. Since the Lab's inception, we have created numerous reference collections in the field of risk prediction and organised a number of challenges based on those datasets. Each challenge centred on a specific risk detection issue, such as depression, anorexia, or self-harm.

eRisk participants only addressed the detection of early risk of depression in its first edition (2017) [6]. This resulted in the first proposals for exploiting the relationship between the use of language in social networks and early signs of depression. Because it was the first edition of such an innovative evaluation scheme, eRisk 2017 was extremely demanding for both participants and organisers. Temporal data chunks were released in sequential order (one chunk per week). Following each release, participants were required to send their predictions about the users in the collection. Only eight of the thirty participating groups completed the tasks by the deadline. More than 30 different approaches to the problem were proposed by these teams (variants or runs). The evaluation methodology and metrics were those defined in [5].

eRisk [7] included two shared tasks in 2018: 1) a continuation of the task on early detection of depression from 2017 and 2) a task on early detection of signs of anorexia. Both tasks were organised similarly and used the same eRisk 2017 evaluation methods. eRisk 2018 had 11 final participants (out of 41 registered). Participants submitted 45 systems for Task 1 (depression) and 35 for Task 2 (anorexia).

In 2019, we organised three tasks [8]. Task 1 was a continuation of the 2018 task on early detection of indicators of anorexia. Task 2 was a new one on early

detection of signs of self-harm. Furthermore, and more novel, a new activity, Task 3, was developed to automatically fill out a depression questionnaire based on user interactions in social media. It should be noted that this new task does not address early detection but rather another complex task (fine-grained depression level estimation). For eRisk 2019, 14 participants (out of 62 registered teams) actively participated in the three challenges and submitted 54, 33, and 33 system versions (runs) for each one, respectively.

We opted to continue the tasks of early detection of self-harm (Task 1) and estimating the severity of depression symptoms (depression level estimation, Task 2) for the 2020 edition [9]. Task 1 had 12 final participants who submitted 46 possible system variants, whilst Task 2 had six active participants who presented 17 different system variants.

Finally, in 2021 we proposed three tasks to the participants. Following our three-year-per-task cycle, we closed the early detection of signs of self-harm challenge (Task 2) and the estimation of the severity of the symptoms of depression (Task 3). Additionally, we presented a new domain for early detection, in this case, pathological gambling (Task 1) [11]. We received 115 runs from 18 teams out of 75 registered. Those are distributed as follows: 26 systems for Task 1, 55 for Task 2 and 34 for Task 3.

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. For celebrating those five years of efforts by participants, we are just finishing the edition of a book on the topic of eRisk [3].

2.1 Early Risk Prediction Tasks

The majority of the proposed challenges were geared toward early risk prediction in various domains (depression, anorexia, self-harm). They were all organised in the same way: the teams had to analyse social media writings (posts or comments) sequentially (in chronological order) to detect risk signals as soon as possible. The main objective was to produce useful algorithms and models for monitoring social network activity.

The social media platform Reddit was used as a source for all shared tasks in the various versions. It is vital to note that Reddit's terms of service allow for data extraction for research purposes. Except as provided by the notion of fair use, Reddit does not enable unauthorised commercial use or redistribution of its material. The research activities of eRisk are an example of fair usage.

Reddit users frequently portray a highly active profile with a large thread of submissions (covering several years). In terms of mental health problems, there are subcommunities (e.g. subreddits) dedicated to depression, anorexia, self-harm, and pathological gambling, to name a few. We used these valuable sources to create collections of writings (posts or comments) made by *redditors* for the eRisk test collections (as mentioned in [5]). In our datasets, *redditors* are divided into positive class (e.g., depressed) and negative class (control group).

When building the datasets, we followed the same approach as Coppersmith and colleagues [2]. We generated the positive class by employing a retrieval strat-

egy for identifying *redditors* who were diagnosed with the disease at hand (e.g. depressed). This was determined through searches for self-expressions associated with medical diagnosis (e.g. "Today, I was diagnosed with depression"). Many *redditors* are active on subreddits about mental health, and they are frequently open about their medical condition. Following that, we carefully checked the collected results to ensure that the expressions about diagnosis were genuine. For example, "*I am anorexic*", "*I have anorexia*", or "*I believe I have anorexia*" were not deemed explicit affirmations of a diagnosis. We only included a user in the positive set where there was a clear and explicit indication of a diagnosis (e.g., "*Last month, I was diagnosed with anorexia nervosa*", "*After struggling with anorexia for a long time, last week I was diagnosed*"). We have a high level of confidence in the reliability of these labels. This semi-automatic extraction method has successfully extracted information about patients who have been diagnosed with a particular disease. Since 2020, we have used Beaver [10], a new tool for labelling positive and negative instances, for aiding us in this task.

The first edition of eRisk presented a new measure called ERDE (Early Risk Detection Error) for measuring early detection [5]. This metric served as a supplement to normal classification measures, which neglect prediction delay. ERDE considers the correctness of the (binary) decision as well as the latency, which is calculated by counting the number (k) of texts seen prior to reaching the decision.

Later on, eRisk added a ranking-based way to evaluate participation. Since 2019, after each round of writings, a user ranking has been generated (ranked by decreasing estimated risk). These ranks are assessed using common information retrieval metrics (for example, P@10 or nDCG). The ranking-based evaluation is described in detail in [8] We have also embraced $F_{latency}$ in 2019, an alternative assessment metric for early risk prediction proposed by Sadeque et al. [12].

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. The Depression Level Estimation Task investigates the feasibility and possible ways for automatically measuring the occurrence and intensity of numerous well-known depression symptoms. In this task, participants had access to the whole history of writings of some *redditors*. With that in hand, participants had to devise models that fill out a standard depression questionnaire using each user's history. Models have to capture evidence from users texts to decide on the answer to each questionnaire item. The questionnaire 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). The questionnaire is based on the Beck's Depression Inventory (BDI) [1].

To produce the ground truth, we compiled a series of surveys filled out by social media users together with their writing history. Because of the unique nature of the task, new evaluation measures for evaluating the participants' estimations were required. We defined four metrics: Average Closeness Rate

(ACR), Average Hit Rate (AHR), Average DODL (ADODL), and Depression Category Hit Rate (DCHR), details can be found in [8].

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 [6–9,11].

So far, we have presented eight editions of early risk tasks on four mental health issues. Participants contributing to those past editions have presented a wide variety of models and approaches. The majority of the methods are based on standard classification techniques. That is, most of our competitors were centred on optimising classification accuracy on the training data. In general, participants were less concerned with the accuracy-delay trade-off. In terms of performance, the results over the years demonstrate some variances between challenges, with anorexia detection yielding better results than depression detection. These differences may be attributable to the amount and quality of the released training data and the very nature of the disorder. We hypothesise that, depending on the condition, patients are more or less prone to leave traces of the language used in social media. The performance figures show a trend on how participants improved detection accuracy edition over edition. This trend motivates us to continue supporting research on text-based early risk screening in social media. Furthermore, given the success of some participants, it appears that automatic or semi-automatic screening systems that predict the commencement of specific hazards are within reach.

In terms of estimating depression levels, the results show that automatic analysis of the user's writings could be a complementary strategy for extracting some signs or symptoms associated with depression. Some participants had a 40 per cent hit rate (the systems answered, i.e., 40% of the BDI questions with the exact same response given by the real user). This still has a lot of room for improvement, but it shows that the participants were able to extract some signals from the chaotic social media data.

The difficulties in locating and adapting measures for these novel jobs has also prompted us to develop new metrics for eRisk.Some eRisk participants [12,13], were also engaged in proposing novel modes of evaluation, which is yet another beneficial outcome of the Lab. We are also planning to incorporate new metrics for automatic risk estimation tasks. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were two widely metrics used in rating prediction for users in recommendation systems [4]. We think that they may be suitable metrics for the problem.

3 eRisk 2022

The results of past editions have inspired us to continue with the Lab in 2022 and further investigate the relationship between text-based screening from social

media and risk prediction and estimation. The scheme of tasks for eRisk 2022 is as follows:

- Task 1: Early detection of pathological gambling. We will continue the new task from 2021. This is the second edition of the task, so following our new three-year cycle, in 2022, participants will have training data.
- Task 2: Early detection of depression. After the second edition in 2018, we will close the cycle for this task with its edition next year. Moreover, and different from previous editions for the disease, the participants will have a post-by-post release of user history through our web service.
- Task 3: Measuring the severity of the signs of eating disorders. This is a new severity estimation task in the field of eating disorders. Eating disorders (ICD-10-CM code F50) affect up to 5% of the population, most often developing in adolescence and young adulthood. The task consists of estimating the severity of the eating disorder from a thread of user submissions. The participants will be given a history of postings for each user, and the participants will have to fill a standard questionnaire (based on the evidence found in the history of postings). The questionnaire that we will use is the Eating Disorder Examination Questionnaire (EDE-Q). 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).

4 Conclusions

The results achieved so far under eRisk and the engagement of the research community motivate us to continue with the proposal of new shared-tasks related to early risk detection. We are truly thankful to participants for their contribution to the success of eRisk. We want to encourage the research teams working on the early risk to keep improving and creating new models for future tasks and risks. Even if generating new resources is tedious, we are convinced that the societal benefits outweigh the costs.

Acknowledgements. This work was supported by projects RTI2018-093336-B-C21, RTI2018-093336-B-C22 (Ministerio de Ciencia e Innvovación & ERDF). The first and second authors thank the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G/01, ED431B 2019/03) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT of the University of A Coruña as a Research Center of the Galician University System. The third author also thanks the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2018/29) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System. 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. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: ACL Workshop on Computational Linguistics and Clinical Psychology (2014)
3. Crestani, F., Losada, D.E., Parapar, J.: Early Detection of Mental Health Disorders by Social Media Monitoring. Springer, Berlin (2021)
4. 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)
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.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
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 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. http://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. 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, pp. 495–503. WSDM 2018, ACM, New York, NY, USA (2018)
13. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. **32**(3), 588–601 (2018)