eRisk 2021: Pathological Gambling, Self-Harm and Depression Challenges

Javier Parapar[0000−0002−5997−8252]1, Patricia Martín-Rodilla[0000−0002−1540−883X]1, David E. Losada[0000−0001−8823−7501]1, and Fabio Crestani[0000−0001−8672−0700]3

[1] Information Retrieval Lab,
Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC),
Universidade da Coruña,
javierparapar@udc.es, patricia.martin.rodilla@udc.es
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
david.losada@usc.es
[3] Faculty of Informatics,
Universitá della Svizzera italiana (USI), Switzerland
fabio.crestani@usi.ch

Abstract. eRisk, a CLEF lab oriented to early risk prediction on the Internet, started in 2017 as a forum to foster experimentation on early risk detection. After four editions (2017, 2018, 2019 and 2020), the lab has created many reference collections in the field and organized multiple early risk detection challenges using those datasets. Each challenge focused on a specific early risk detection problem (e.g., depression, anorexia or self-harm). This paper describes the work done so far, discusses the main lessons learned over the past editions and the plans for the eRisk 2021 edition, where we introduced pathological gambling as a new early risk detection challenge.

1 Introduction

As a part of CLEF (Conference and Labs of the Evaluation Forum), the eRisk lab is a forum for exploring the evaluation methodology and effectiveness metrics related to early risk detection on the Internet (with past challenges particularly focused on health and safety). Over the past editions [8,7,6,5], a number of testbeds and tools have been developed under the eRisk's umbrella. eRisk's dataset building methodology and the evaluation strategies proposed are general and, thus, potentially applicable to multiple application domains.

This lab brings together different research disciplines (e.g. information retrieval, computational linguistics, machine learning or psychology) to address the posed problems in an interdisciplinary way. Furthermore, effective solutions to eRisk tasks are potentially applicable to socially important concerns. For example, systems may send warning alerts when an individual starts broadcasting suicidal thoughts or threats of self-harm on Social Media. Previous editions of

eRisk proposed shared tasks focused on specific health and security problems, such as depression, anorexia or self-harm detection.

eRisk takes an iterative approach, where risk prediction is seen as a sequential process of accumulation of evidence. The constant production of data in a given data source (e.g. Social Media entries) needs to be automatically analyzed by the systems designed by eRisk participants. Within this process, the algorithms need to estimate when and if there is enough aggregated evidence about a certain type of risk. The shared tasks represent a successful methodology for improving results collaboratively about different types of risks. On each shared task, the participants have access to a temporally organized dataset where they have to balance between making *early* alerts (e.g., based on few social media entries) or *not-so-early* (late) alerts (e.g., evaluating a wider range of entries and only emit alerts after analyzing a larger number of pieces of evidence).

2 Previous Editions of eRisk

eRisk, a CLEF lab for research on early risk prediction on the Internet, started in 2017 as a forum to set the experimental foundations of early risk detection. After four editions (2017, 2018, 2019 and 2020), the lab has created many reference collections in the field and organized several early risk detection challenges using those datasets. Each challenge focused on a specific early risk detection problem, such as depression, anorexia and self-harm.

In the first edition (2017) [5], eRisk focused on the detection of early signs of depression, trying to explore the relationship between the use of language in social networks and early signs of depression. It was the first edition of such an innovative evaluation scheme and, thus, eRisk 2017 was very demanding for both the participants and the organizers. Temporal data chunks were released sequentially (one chunk per week). After each release, the participants had to send their predictions about the users in the collection. Only 8 of the 30 participating groups completed the tasks by the required deadline. These teams proposed more than 30 different interdisciplinary approaches to the problem (variants or runs). The evaluation methodology and metrics were those defined in [4].

In 2018, eRisk [6] included two shared tasks: 1) a continuation of 2017's task on early detection of depression and 2) a task on early detection of signs of anorexia. Both tasks followed a similar organization and the same evaluation methods of eRisk 2017. eRisk 2018 had 11 final participants (out of 41 registered), proposing 45 runs for Task 1 and 35 runs for Task 2.

In 2019, we organized three tasks [7], Task 1 as a continuation of 2018 task on early detection of signs of anorexia and Task 2, a new one on early detection of signs of self-harm. Furthermore, a new task, Task 3, was introduced oriented to automatically filling a depression questionnaire based on user interactions in social media. Note that Task 3 does not address early detection but another complex task (depression level estimation). For eRisk 2019, 14 participants (out of 62 registered teams) actively participated in the three tasks and submitted 54, 33 and 33 system variants (runs), respectively for each task.

Finally, the last edition of eRisk (2020) [8] continued the task of early detection of self-harm (task 1) and the task of measuring the severity of the signs of depression (depression level estimation, task 2). Task 1 had 12 final participants who submitted 46 different variants, while task 2 had six active participants who proposed 17 different system variants (runs).

Over these four years, eRisk has received a steady number of active participants, slowly placing the lab as a reference forum for early risk research.

2.1 Early Risk Prediction Tasks

Most of the proposed shared tasks were oriented to the early prediction of risk in different challenges (depression, anorexia, self-harm) whereas one specific task addresses the estimation of the level of depression.

Regarding the former group of tasks, all of them followed the same organization: the teams had to sequentially (following chronological order) process social media writings –posts or comments– intending to detect signs of risk as soon as possible. The resulting algorithms represent effective solutions for monitoring social network activity. A summary of the main statistics of the collections used in the early risk detection task over the years is shown in Table 1.

Reddit was the social media platform used as a source for all shared tasks in the different editions. It is important to highlight that Reddit's terms of use permit to extract data for research purposes. Reddit does not permit the unauthorized commercial use of its contents or redistribution, except as permitted by the doctrine of fair use. eRisk's research activities are an example of fair use.

Commonly, users in Reddit present a highly active profile, with a large thread of submissions (covering several years). Regarding psychological disorders, there are specific subcommunities (*subreddits*) about depression, anorexia, and self-harm, just to name a few. We used these valuable sources for building the eRisk test collections (as we described in [4].), creating collections of writings (posts or comments) published by *redditors*. Redditors are classified into two classes: the positive class (e.g., depressed) and the negative class (control group).

Following the method proposed by Coppersmith and colleagues [3], the positive class was obtained using a retrieval approach for identifying *redditors* diagnosed with the condition at hand (e.g. depressed). This was based on searches for self-expressions related to medical diagnoses (e.g. "Today, I was diagnosed with depression"). Many *redditors* are active on subreddits related to psychological disorders and, often, they tend to be very explicit about their medical condition. Next, we manually reviewed the retrieved results to verify that the expressions about diagnosis look really genuine. For example, expressions such as "I am anorexic", "I have anorexia" or "I think I have anorexia" were not considered as explicit expressions of a diagnosis. We only included a user into the positive set when there was a mention of a diagnosis that was clear and explicit (e.g., "Last month, I was diagnosed with anorexia nervosa", "After struggling with anorexia for a long time, last week I was diagnosed"). Our confidence in the reliability of these labels is high. This semi-automatic extraction method

has been successful in retrieving information about people diagnosed with a specific disorder. In 2020, we introduced the use of Beaver, a new tool for labelling positive and negative cases [9].

For evaluating early detection, the first editions of eRisk considered a new measure called ERDE (Early Risk Detection Error) [4]. This measure acted as a complement of standard classification metrics, which ignore the delay in making predictions. ERDE takes into account the correctness of the (binary) decision and the delay, which is measured by counting the number (k) of writings seen before making the decision. From the 2019 edition, eRisk also incorporated a ranking-based approach to evaluate the participants: a user ranking was produced after each round of writings (ranked by decreasing estimated risk) and these rankings were evaluated under standard information retrieval metrics (e.g., P@10 or NDCG). The ranking-based evaluation is fully detailed in [7]. Since eRisk 2019, we also adopted $F_{latency}$, an alternative evaluation metric for early risk prediction that was proposed by Sadeque and colleagues [10].

2.2 Severity Level Estimation Task

One specific task in 2019 and 2020 was dedicated to estimating the severity level of depression. Depression Level Estimation Task explores the viability and possible approaches for automatically estimating the occurrence and intensity of multiple well-known symptoms of depression. In these tasks, the participants had access to the full history of writings of a number of redditors, and each group had to design an automatic method that reads the history of each user and fills a standard depression questionnaire based on the evidence found in the user's writings. The questionnaire included 21 questions (with four possible responses corresponding with different severity levels) about the intensity of depression signals and symptoms (e.g., loss of energy, sadness, and sleeping problems). The questionnaire is derived from the Beck's Depression Inventory (BDI) [2].

The ground truth for this task was a collection of questionnaires directly filled by social media users, together with their history of writings. Due to the specific nature of the task, it was necessary to introduce evaluation metrics for evaluating the participants' estimations. We considered four metrics [7]: Average Closeness Rate (ACR), Average Hit Rate (AHR), Average DODL (ADODL) and Depression Category Hit Rate (DCHR).

2.3 Results

Yearly reports with a full description and critical analysis of eRisk results have been published since 2017 [8,7,6,5]. The early risk prediction tasks have involved a wide range of participants and variants. Most of the approaches are based on traditional classification workflows (centred on obtaining effective classifiers from the training data). In general, the participants paid less attention to the accuracy-delay tradeoff. In terms of performance, the results show some differences between challenges, with, for example, more effective results in anorexia

	Training Stage		*Test Stage*	
	eRisk 2017 - Depression Task			
	Depressed	*Control*	*Depressed*	*Control*
Num. subjects	83	403	52	349
Num. submissions (posts & comments)	30,851	264,172	18,706	217,665
Avg num. of submissions per subject	371.7	655.5	359.7	623.7
Avg num. of days from first to last submission	572.7	626.6	608.31	623.2
Avg num. words per submission	27.6	21.3	26.9	22.5
	eRisk 2018 - Depression Task			
	Depressed	*Control*	*Depressed*	*Control*
Num. subjects	135	752	79	741
Num. submissions (posts & comments)	49,557	481,837	40,665	504,523
Avg num. of submissions per subject	367.1	640.7	514.7	680.9
Avg num. of days from first to last submission	586.43	625.0	786.9	702.5
Avg num. words per submission	27.4	21.8	27.6	23.7
	eRisk 2018 - Anorexia Task			
	Anorexia	*Control*	*Anorexia*	*Control*
Num. subjects	20	132	41	279
Num. submissions (posts & comments)	7,452	77,514	17,422	151,364
Avg num. of submissions per subject	372.6	587.2	424.9	542.5
Avg num. of days from first to last submission	803.3	641.5	798.9	670.6
Avg num. words per submission	41.2	20.9	35.7	20.9
	eRisk 2019 - Anorexia Task			
	Anorexia	*Control*	*Anorexia*	*Control*
Num. subjects	61	411	73	742
Num. submissions (posts & comments)	24,874	228,878	17,619	552,890
Avg num. of submissions per subject	407.8	556.9	241.4	745.1
Avg num. of days from first to last submission	$\approx$ 800	$\approx$ 650	$\approx$ 510	$\approx$ 930
Avg num. words per submission	37.3	20.9	37.2	21.7
	eRisk 2019 - Self-harm Task			
	Self-harm	*Control*	*Self-harm*	*Control*
Num. subjects	-	-	41	299
Num. submissions (posts & comments)	-	-	6,927	163,506
Avg num. of submissions per subject	-	-	169.0	546.8
Avg num. of days from first to last submission	-	-	$\approx$ 495	$\approx$ 500
Avg num. words per submission	-	-	24.8	18.8
	eRisk 2020 - Self-harm Task			
	Self-harm	*Control*	*Self-harm*	*Control*
Num. subjects	41	299	104	319
Num. submissions (posts & comments)	6,927	163,506	11,691	91,136
Avg num. of submissions per subject	169.0	546.8	112.4	285.6
Avg num. of days from first to last submission	$\approx$ 495	$\approx$ 500	$\approx$ 270	$\approx$ 426
Avg num. words per submission	24.8	18.8	21.4	11.9

Table 1. Statistics of the train and test collections used in the early prediction tasks.

detection than those in depression. The performance figures showed how participants managed to improve the detection accuracy edition by edition. This encourages us to keep fostering research on text-based early risk screening from social media. Furthermore, given the effectiveness achieved by some participants, it appears that automatic or semi-automatic screening tools that estimate the onset of certain risks are within reach.

The difficulty in finding and adjusting metrics for these innovative tasks has also motivated us to incorporate new metrics for eRisk. Some eRisk participants [10,11] were also active in proposing new forms of evaluation, which is another valuable result of the lab.

Regarding depression level estimation, the results suggest that automatic analysis of the user's writings might be a complementary approach for extracting

some signals or symptoms related to depression. Some participants had a hit rate of 40% (i.e., 40% of the BDI questions were answered by the systems with the exact same response given by the real user). This has still much room for improvement, but, in any case, it suggests that the participants were able to extract some signal from the noisy Social Media data.

3 Conclusions and Future Work

The results achieved so far encourage us to continue with the lab in 2021 and further explore the relation between text-based screening from social media and early risk. For eRisk 2021, our plan is twofold:

- Firstly, expanding the range of target domains for early risk detection from social networks. Specifically, eRisk 2021 presents as Task 1 the early detection of risks in pathological gambling, a growing psychological disorder. Pathological gambling (ICD-10-CM code F63.0) is also called ludomania and usually referred to as *gambling addiction* (it is an urge to gamble independently of its negative consequences). According to the World Health Organization [1], in 2017, adult gambling addiction had prevalence rates ranged from 0.1% to 6.0%. Following our usual methodology, we will collect and release data in a sequential way. The participating systems will interact with a server prepared for this task in order to collect data and send results.
- Secondly, we will establish an (at least) three year cycle per task, where we will not release training data in the first year (as it happened in the first edition of self-harm). The objective is to foster research on methods that do not solely depend on the existence of training. Then, in the second edition, we will see how the performance of the systems can be improved with training data. Finally, in the third edition, we will see how participants manage to improve and refine their models after two years of experience.
- Following the scheme suggested above, in 2021, we present the third edition of two already existing tasks: a shared task will be organized on early detection of self-harm (2021's Task 2), and a task on estimating the severity of the signs of depression (2021's Task 3, based on standard depression questionnaire).

Acknowledgements

We thank the financial support obtained from the i) "Ministerio de Ciencia, Innovación y Universidades" of the Government of Spain (research grants RTI2018-093336-B-C21 and RTI2018-093336-B-C22), ii) "Consellería de Educación, Universidade e Formación Profesional" , Xunta de Galicia (grants ED431G 2019/01 and ED431G 2019/04). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

1. Abbott, M.: The epidemiology and impact of gambling disorder and other gambling-related harm. In: WHO Forum on alcohol, drugs and addictive behaviours. Geneva, Switzerland (2017)
2. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An Inventory for Measuring Depression. JAMA Psychiatry **4**(6), 561–571 (06 1961)
3. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: ACL Workshop on Computational Linguistics and Clinical Psychology (2014)
4. 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)
5. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 346–360. Springer International Publishing, Cham (2017)
6. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: Early risk prediction on the internet. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 343–361. Springer International Publishing, Cham (2018)
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 early risk prediction on the internet. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 340–357. Springer International Publishing, Cham (2019)
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2020: Early risk prediction on the internet. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 272–287. Springer International Publishing, Cham (2020)
9. 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, July 6-9, 2020 (2020), `http://ceur-ws.org/Vol-2621/CIRCLE20_23.pdf`
10. 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 '18, ACM, New York, NY, USA (2018)
11. Trotzek, M., Koitka, S., Friedrich, C.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering (04 2018)