eRisk 2020: Self-harm and Depression Challenges

David E. Losada[1(✉)], Fabio Crestani[2], and Javier Parapar[3]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
david.losada@usc.es

[2] Faculty of Informatics, Universitá della Svizzera italiana (USI), Lugano, Switzerland
fabio.crestani@usi.ch

[3] Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña, A Coruña, Spain
javierparapar@udc.es

Abstract. This paper describes eRisk, the CLEF lab on early risk prediction on the Internet. eRisk started in 2017 as an attempt to set the experimental foundations of early risk detection. Over the last three editions of eRisk (2017, 2018 and 2019), the lab organized a number of early risk detection challenges oriented to the problems of detecting depression, anorexia and self-harm. We review in this paper the main lessons learned from the past and we discuss our future plans for the 2020 edition.

1 Introduction

eRisk is a CLEF lab whose main goal is to explore issues of evaluation methodology, performance metrics and other challenges related to building testbeds for early risk detection [4–6]. The predictive tools developed under eRisk's shared tasks could be potentially useful in different areas, particularly those related to health and safety. For example, warning alerts can be sent when an individual starts broadcasting suicidal thoughts on Social Media. eRisk tries to instigate interdisciplinary research (e.g. related to information retrieval, machine learning, psychology, and computational linguistics) and the advances developed under this challenge would be potentially applicable to support a number of socially important problems.

The lab casts early risk prediction as a process of *sequential accumulation of evidence*. In other words, given a stream of data (e.g. real-time Social Media entries), alerts should be fired when there is enough evidence about a certain type of risk. The participants have access to a stream of social media entries and they have to balance between making *early* alerts (e.g., based on few entries or posts) or *not-so-early* (late) alerts (e.g., if participants opt to see a wider range of entries and only emit alerts after thoroughly analyzing the available pieces of evidence). The testset building methodology and the evaluation strategies proposed

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 557–563, 2020.
https://doi.org/10.1007/978-3-030-45442-5_72

under eRisk are general and, thus, potentially applicable to multiple application domains (for example, health, security, or cybergrooming). However, all previous eRisks have focused on tasks and data related to psychological disorders.

2 Previous Editions of eRisk

eRisk 2017 included an exploratory task on early detection of signs of depression. This shared task was defined using the test collection and evaluation metrics proposed in [3]. The interactions between depression and natural language use is an intriguing problem and the eRisk participants approached the challenge from multiple perspectives.

The 2017 task was demanding both for the participants and the organisers because it had ten releases of data, and, after each release, the teams had one week to submit their predictions. Furthermore, eRisk was new to all participants and the research groups were not familiar with this novel evaluation metrics. As a result, only eight teams (out of 30 registered participants) were able to follow the tight schedule, submitting thirty different system variants (or runs).

In 2018, two shared tasks were organized: task 1, on early detection of signs of depression (which was a continuation of the 2017's pilot task), and task 2 on early detection of signs of anorexia. The two tasks had the same overall organization and evaluation method of the previous year. eRisk 2018 had 11 active participants (out of 41 registered teams) who submitted 45 and 35 system variants (for Task 1 and Task 2, respectively).

In 2019, three shared tasks were organized. Two of them were oriented to the same early detection technologies (task 1, early detection of signs of anorexia; task 2, early detection of signs of self-harm), while a new task (task 3) was introduced oriented to automatically filling a depression questionnaire based on user interactions in social media. eRisk 2019 had 14 active participants (out of 62 registered teams) who submitted 54, 33 and 33 system variants (respectively, for the 3 tasks). The increasing numbers of participants and runs submitted suggest that eRisk is slowly becoming an experimental reference for early risk research.

2.1 Early Risk Prediction Tasks

Previous eRisk's early prediction tasks consisted of sequentially processing writings –posts or comments– published by social media users and learn to detect signs of risk as soon as possible. The participating systems had to process the writings in chronological order (oldest writings are given first to the participants). In this way, algorithms that effectively perform this shared task could be applied to monitor interactions in blogs, social networks, or other types of Social Media. Table 1 reports the main statistics of the collections utilized in the early prediction tasks of eRisk 2017–2019.

Reddit was the source of data for these shared tasks. It is a social media platform where users (*redditors*) post and vote submissions which are organized by communities of interests (*subreddits*). Reddit has a large set of users and many

of them have a large thread of submissions (covering several years). Reddit has active subreddits about psychological disorders, such as depression or eating disorders. Reddit's terms and conditions permit to use its contents for research purposes[1].

The test collections used in the eRisk early detection tasks have the same format as the collection described in [3]. It is a collection of writings (posts or comments) published by redditors. For each task, there are two classes of redditors: the positive class (e.g., depression or anorexia) and the negative class (control group). The positive class was obtained following the extraction method proposed by Coppersmith and colleagues [2] (an automatic approach to identify people diagnosed with depression in Twitter). We adapted this extraction approach to Reddit as follows. Self-expressions related to medical diagnoses (e.g. "Today, I was diagnosed with depression") can be obtained by running specific phrases against the platform search tool. Next, we manually reviewed the retrieved results to verify that they were really genuine. Our confidence on the reliability of these labels is high. There are many subreddits oriented to people suffering from psychological disorders and, usually, many redditors are active on these subreddits. These users tend to be very explicit about their problems and medical condition. This extraction approach is semi-automatic (requires manual revision of the retrieved posts), but it is an effective way to extract a group of people that are diagnosed with a given disorder. The manual reviews were thorough and strict. Expressions such as "I have anorexia", "I think I have anorexia", or "I am anorexic" 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"). For each redditor, the test collection contains his sequence of writings (in chronological order) and each task was organized into a **Training stage**, where the participants had access to training data (we released the full history of writings published by a set of training redditors), and a **Test stage**. In eRisk 2017 and eRisk 2018, the test stage was organized in a 10-week format as follows. The sequence of writings published by each user was split into 10 chunks (the first chunk has the oldest 10% of the user's writings, the second chunk has the second oldest 10%, and so forth). The test stage had 10 releases of data (one release per week). The first week we gave the first chunk of data to the participants, the second week we gave the second chunk of data, and so forth. After each release, the participants had to process the data and, before the next release, each participant had to choose between: (a) emitting a decision on the redditor (positive or negative), or (b) making no decision (i.e. waiting to see more chunks). This choice had to be made for each redditor in the test set.

[1] Reddit privacy policy states explicitly that the submitted posts and comments are not private and will still be accessible after the user's account is deleted. Reddit does not permit unauthorized commercial use of its contents or redistribution, except as permitted by the doctrine of fair use. These research activities are an example of fair use.

If the participant emitted a decision then the decision was considered as final. The systems were evaluated based on the accuracy of the decisions and the number of chunks required to take the decisions (see below). In 2019, we moved from this "chunk-based" release of test data to a "item-by-item" release of test data. We set up a REST server that iteratively provided the user's writings to the participants[2]. In this way, each participant could stop and make an alert at any point of the user's chronology (the server waited for the responses of the participants and only gave new user data after receiving the participants' input).

Evaluation Metrics for Early Risk Detection. The evaluation of these tasks considered standard classification measures, such as F1, Precision and Recall, computed with respect to the positive group. These standard classification measures evaluate the participants' estimations with respect to golden truth labels. eRisk included them in the evaluation reports because these measures are well-known and interpretable. However, these three measures are time-unaware and, thus, do not penalize late alerts. In order to reward early detection algorithms, we introduced in [3] a new measure called ERDE (Early Risk Detection Error). 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.

In eRisk 2019 the set of evaluation metrics was extended. We complemented the evaluation report with additional decision-based metrics that try to capture additional aspects of the problem. We adopted $F_{latency}$, an alternative evaluation metric for early risk prediction that was proposed by Sadeque and colleagues [7]. Another novelty introduced in 2019 was that user's data was processed by the participants in a post by post basis (as opposed to the old chunk-based approach). Besides decision-based evaluation metrics, eRisk 2019 incorporated a ranking-based approach to evaluate the participants. This form of evaluation was based on rankings of users by decreasing estimated risk. These rankings were produced after each round of writings and were evaluated with standard information retrieval measures, such as P@10 or NDCG. A full description of this ranking-based evaluation approach can be found in the eRisk 2019's overview report [6].

2.2 Depression Level Estimation Task

Introduced in 2019, the task consisted of estimating the level of depression from a thread of user submissions. For each user, the participants were given a full history of writings (in a single release of data) and the participants had to fill a standard depression questionnaire based on the evidence found in the history of postings. The questionnaire is derived from the Beck's Depression Inventory (BDI) [1], which assesses the presence of feelings like sadness, pessimism, loss of

[2] More information about the server and the modality of release of the date can be found at the eRisk's website on http://early.irlab.org/server.html.

Table 1. Main statistics of the train and test collections used in the early prediction tasks of eRisk 2017–2019.

	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	≈800	≈650	≈510	≈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	–	–	≈495	≈500
Avg num. words per submission	–	–	24.8	18.8

energy, etc. for the detection of depression. The questionnaire contains 21 questions and each question has a set of at least four possible responses, ranging in intensity. For example, the question on sadness has these four possible responses: (0) I do not feel sad, (1) I feel sad, (2) I am sad all the time and I can't snap out of it, and (3) I am so sad or unhappy that I can't stand it.

The task aimed at exploring the viability of automatically estimating the severity of the multiple symptoms associated with depression. Given the user's

history of writings, the algorithms had to estimate the user's response to each individual question. We collected questionnaires filled by social media users together with their history of writings (we extracted each history of writings right after the user provided us with the filled questionnaire). The questionnaires filled by the users (ground truth) were used to assess the quality of the responses provided by the participants.

Four evaluation measures were introduced to evaluate the participants' estimations. The Average Hit Rate (AHR) computes the ratio of cases where the *automatic questionnaire* has exactly the same answer as the real questionnaire. The Average Closeness Rate (ACR) is a less stringent measure that considers the distance between each real answer and the answer submitted by the participating team. The two other measures, ADODL and DHCR, were oriented to compute how effective the systems are at estimating the overall depression level of the individual. These two measures compute the deviation between the total depression score (sum of all responses in the questionnaire) of the real questionnaire vs the questionnaire submitted by the participants.

2.3 Results

A full description and analysis of the results can be found in the lab overviews [4–6] and working note proceedings. For the early risk prediction tasks, most of the participating teams focused on classification aspects (i.e. how to learn effective classifiers from the training data) and no much attention was paid to the tradeoff between accuracy and delay. For the depression level estimation task, the results show that an automatic analysis of the user's writings is useful at extracting some signals or symptoms related to depression (e.g., some participant had a hit rate of 40%).

Although the effectiveness of the proposed solutions is still modest, the experiments performed under these shared tasks suggest that evidence extracted from social media is valuable. Automatic or semi-automatic screening tools are indeed promising to detect at-risk individuals. This result encouraged us to continue with the lab in 2020 and further explore the creation of new benchmarks for text-based screening of signs of such risks.

Another important outcome of the previous eRisk labs is related to the evaluation methodology. How to define appropriate metrics for early risk prediction is a challenge by itself and eRisk labs have already instigated the development of new early prediction metrics [7,8].

3 Conclusions and Future Work

eRisk will continue at CLEF 2020. Our plan is to organize two shared tasks. The first task will be a continuation of 2019's eRisk task on early detection of signs of self-harm. The second task will be a continuation of 2019's task on depression level estimation. In 2019, these two tasks were really challenging and the participants had no training data. In 2020, we will use the eRisk 2019 data

as training data, and new test cases will be collected and included into the 2020 test split. By running these two tasks again we expect to further gain insight into the main factors and issues related to extracting signs of self-harm and depression from Social Media entries.

Acknowledgements. We thank the support obtained from the Swiss National Science Foundation (SNSF) under the project "Early risk prediction on the Internet: an evaluation corpus", 2015. We also 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 ED431C 2018/29, ED431G/08 and ED431G/01 – "Centro singular de investigación de Galicia" –). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

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. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 28–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_3
4. 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
5. 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
6. 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
7. 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)
8. 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**(3), 588–601 (2018)