Early detection of risks on the Internet: an exploratory campaign

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

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

Abstract. This paper summarizes the activities related to the CLEF lab on early risk prediction on the Internet (eRisk). eRisk was initiated in 2017 as an attempt to set the experimental foundations of early risk detection. The first edition essentially focused on a pilot task on early detection of signs of depression. In 2018, the lab was enlarged and included an additional task oriented to early detection of signs of anorexia. We review here the main lessons learned and we discuss our plans for 2019.

1 Introduction

The main goal of eRisk, a CLEF lab on early risk detection [4, 3], is to explore issues of evaluation methodology, performance metrics and other issues related to building testbeds for early risk detection. Early risk detection can be useful in different areas, particularly those related to health and safety. For instance, warning alerts can be given when a predator starts contacting a child for sexual purposes, or when an offender publishes antisocial threats on Social Media. eRisk intends to pioneer a new interdisciplinary area of research whose results would be potentially applicable to detect potential paedophiles, stalkers, individuals with a latent tendency to fall into the hands of criminal organizations, people with suicidal inclinations, or people susceptible to depression.

The lab views early risk prediction as a process of accumulation of evidence where alerts should be made when there is enough evidence about a certain type of risk. For example, the pieces of evidence could be Social Media posts submitted at various times. A common characteristic of the campaigns run so far is that the pilot tasks worked with stream data and the participating teams had to find a balance between emitting early decisions (based on just a few pieces of evidence) and emitting *not-so-early* decisions (if they opt to wait and analyze more pieces of evidence). Although the collection building strategies and performance metrics are generic and potentially applicable to the usage scenarios described above, all previous editions of eRisk have focused on data related to psychological disorders.

The rest of the paper discusses previous results and sketches our plans for eRisk 2019.

2 Previous editions of eRisk

eRisk 2017 included a pilot task on early detection of depression. This exploratory task was based on the test collection and metrics presented in [2]. The interactions between depression and natural language use is a challenging problem and by sharing this collection with other teams we expected to instigate fruitful discussions on these issues. The 2017 participants employed a wide range of techniques in information access and closely related fields, such as Natural Language Processing, Machine Learning, and Information Retrieval. This pilot task was moderately successful. More than 30 teams registered for this task and got access to the data. The challenge was demanding because it involved ten different releases of data, and, after each release, the teams had one week to submit their results. Furthermore, eRisk was new to all CLEF participants and, thus, the teams were not familiar with the task. As a result, only 8 teams were able to follow this tight and novel schedule. We got 30 different contributions (system variants) from the 8 contributing teams.

In 2018, we proposed two campaign-like tasks: task 1 was a continuation of the pilot task that ran in 2017 (early detection of signs of depression) and task 2 was new (early detection of signs of anorexia). Both tasks had the same structure and evaluation design. Compared with eRisk 2017, eRisk 2018 received increased attention. In 2018, we had 41 registered participants. We got 45 submissions (system variants) for Task 1 and 35 submissions (system variants) for Task 2. There were 11 active teams that engaged into the eRisk tasks. These numbers suggest that the lab is slowly becoming an experimental reference for early detection technologies. In 2019, we expect to increase participation because many groups are now familiar with eRisk and its tight schedule, and some of them have already worked with the data (although they could not make it to send the required results in the previous years).

2.1 Tasks

The tasks consisted of sequentially processing pieces of evidence –in the form of writings (post or comments) posted by Social Media users– and learn to detect early signs of risk as soon as possible. Texts had to be processed by the participating systems in the order they were created. In this way, systems that effectively perform this task could be employed to sequentially track user publications in blogs, social networks, or other types of online media. Table 1 reports the main statistics of the collections utilized in eRisk 2017 and eRisk 2018.

Reddit was the main source of data for our experimental tasks. It is an open-source platform where community members submit content, vote submissions, and publications are organized by areas of interests (*subreddits*). Reddit has a large community of members (*redditors*) and many of the members have a large history of previous submissions (covering several years). It also contains substantive contents about different

medical conditions, such as depression or eating disorders. Reddit's terms and conditions allow to use its contents for research purposes[1].

The test collections used in eRisk 2017 and eRisk 2018 have the same format as the collection described in [2]. It is a collection of publications (posts or comments) done by redditors. For each task, there were two classes of users: the positive class (depression or anorexia, respectively) and a negative class (control group). The positive class was extracted following the approach proposed by Coppersmith et al. [1]. These authors proposed an automatic method to identify people diagnosed with depression in Twitter. We have adapted this estimation method to Reddit as follows. Self-expressions related to diagnoses can be obtained by running specific searches against Reddit (e.g. "I was diagnosed with anorexia"). Next, we manually reviewed the matched posts to verify that they were really genuine. Our confidence on the quality of these assessments is high. In Reddit, there are many support communitities for people suffering from different disorders and it is often the case that redditors go there and are very explicit about their problems and medical condition. Although this method requires manual intervention, it is a simple and effective way to extract a large group of people that explicitly declare having being diagnosed with a given disorder. The manual reviews were strict. Expressions like "I have anorexia", "I think I have anorexia", or "I am anorexic" did not qualify as explicit expressions of a diagnosis. We only included a redditor into the positive group when there was a clear and explicit mention of a diagnosis (e.g., "In 2013, I was diagnosed with anorexia nervosa", "After struggling with anorexia for many years, yesterday I was diagnosed").

For each user, the collection contains his sequence of submissions (in chronological order) and this sequence was split into 10 chunks. The first chunk has the oldest 10% of the submissions, the second chunk has the second oldest 10%, and so forth. Each task was organized into two different stages:

- **Training stage**. Initially, the teams that participated in the task had access to some training data. In this stage, we released the entire history of submissions done by a set of training users. All chunks of all training users were sent to the participants and the actual class of each training user was provided.
- **Test stage**. The test stage had 10 releases of data (one release per week). The first week we gave the 1st chunk of data to the teams (oldest submissions of all test users), the second week we gave the 2nd chunk of data (second oldest submissions of all test users), and so forth. After each release, the teams had to process the data and, before the next week, each team had to choose between: a) emitting a decision on the user (i.e. positive or negative), or b) making no decision (i.e. waiting to see more chunks). This choice had to be made for each user in the test split. If the team 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).

[1] Reddit privacy policy states explicitly that the posts and comments redditors make are not private and will still be accessible after the redditor'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. This research is an example of fair use.

	Train		*Test*	
	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

Table 1. Main statistics of the train and test collections used in the eRisk 2017 and 2018 tasks (depression and anorexia).

2.2 Evaluation metrics for early risk detection

The evaluation of the tasks considered standard classification metrics, such as F1, Precision and Recall (computed with respect to the positive class) and the early risk detection measure proposed in [2]. The standard classification measures evaluate the teams' estimations with respect to golden truth judgments. We included them in our experimental evaluation because these metrics are well-known and easily interpretable. However, they are time-unaware and do not penalize late decisions. In order to reward early alerts, we employed ERDE, an error measure for early risk detection [2] for which the fewer writings required to make the alert, the better.

ERDE (*early risk detection error*) takes into account the correctness of the (binary) decision and the delay, which is measured by counting the number (k) of distinct submissions (posts or comments) seen before taking the decision. For instance, imagine a user u who posted a total number of 150 posts or comments (15 submissions per chunk). If a team's system emitted a decision for user u after the third chunk of data then the delay k would be 45.

Another important factor is that data are unbalanced (many more negative cases than positive cases) and, thus, the evaluation measure needs to weight different errors in a different way. Consider a binary decision d taken by a team's system with delay k. Given golden truth judgments, the prediction d can be a true positive (TP), true negative (TN), false positive (FP) or false negative (FN). Given these four cases, the ERDE measure is defined as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

How to set c_{fp} and c_{fn} depends on the application domain and the implications of FP and FN decisions. In evaluating the systems, we fixed c_{fn} to 1 and c_{fp} was set according to the proportion of positive cases in 2017's test data (e.g. we set c_{fp} to 0.1296). The factor $lc_o(k)(\in [0, 1])$ represents a cost associated to the delay in detecting true positives. We set c_{tp} to c_{fn} (i.e. c_{tp} was set to 1) because late detection can have severe consequences (as a late detection is considered as equivalent to not detecting the case at all). The function $lc_o(k)$ is a monotonically increasing function of k (sigmoid). The latency cost factor was only used for the true positives because we understand that late detection is not an issue for true negatives. True negatives are non-risk cases that, of course, would not demand early intervention (i.e. these cases just need to be effectively filtered out from the positive cases). The systems must therefore focus on early detecting risk cases and detecting non-risk cases (regardless of when these non-risk cases are detected).

2.3 Performance results

In general, the effectiveness of the submitted systems was weak (particularly, for the depression task, see Table 2, which reports the results obtained in the last edition of eRisk). This suggests that the depression task is really challenging and we still need further research on the intriguing aspects of early risk detection. Most of the 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. Only a couple of teams tried to define temporal models that incorporate some sort of sophisticated estimation of the evolution of the disorders. A full description and analysis of the results can be found in the lab overviews [4, 3] and working note proceedings. Another important outcome of eRisk 2017 and eRisk 2018 is related to the evaluation measures. 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 [5, 6].

3 Conclusions and Future Work

eRisk will continue at CLEF 2019. Our plan is to organize up to three different tasks. The first task will be a continuation of 2018's eRisk task on early detection of signs of anorexia. We will use the eRisk 2018 data as training data, and new anorexia and non-anorexia test cases will be collected and included into the 2019 test split. The second task will follow a slightly different format. First, we will provide no training data. In this way, the participants will be encouraged to design predictive methods (e.g. based on search) that require no labelled examples. Most of the systems implemented for the previous eRisk editions were heavily dependent on supervised learning techniques. In 2019, we want to explore some tasks where training data are not available. Second, this task will focus on self-harm problems and, for each individual, the algorithms would be given only the history of the postings *before* the individual entered into the self-harm community. An individual who is active on a self-harm community perhaps has already done some sort of self-harm to his body. We want algorithms that detect the cases earlier

on (and not when the cases are explicit and the individual is already engaging in a support forum). As a consequence, the participants would only be given the texts posted by the affected individuals before they first engaged in the self-harm community (before their first post in this community).

eRisk 2019 will also include a third task on searching for signs of depression. We have collected new data on depression that will consist not only on postings submitted by the depressed users but also on standard questionnaires that estimate their level of depression. Participants will be asked to automatically fill the depression questionnaire based on the user's postings. In this way, we can evaluate how good the algorithms are at detecting multiple elements or symptoms associated with depression.

	depression					anorexia				
	$ERDE_5$	$ERDE_{50}$	F1	P	R	$ERDE_5$	$ERDE_{50}$	F1	P	R
FHDO-BCSGA	9.21%	6.68%	0.61	0.56	0.67	12.17%	7.98%	0.71	0.67	0.76
FHDO-BCSGB	9.50%	**6.44%**	**0.64**	0.64	0.65	11.75%	6.84%	0.81	0.84	0.78
FHDO-BCSGC	9.58%	6.96%	0.51	0.42	0.66	13.63%	9.64%	0.55	0.47	0.66
FHDO-BCSGD	9.46%	7.08%	0.54	0.64	0.47	12.15%	**5.96%**	0.81	0.75	**0.88**
FHDO-BCSGE	9.52%	6.49%	0.53	0.42	0.72	11.98%	6.61%	**0.85**	0.87	0.83
LIIRA	9.46%	7.56%	0.50	0.61	0.42	12.78%	10.47%	0.71	0.81	0.63
LIIRB	10.03%	7.09%	0.48	0.38	0.67	13.05%	10.33%	0.76	0.79	0.73
LIIRC	10.51%	7.71%	0.42	0.31	0.66					
LIIRD	10.52%	7.84%	0.42	0.31	0.66					
LIIRE	9.78%	7.91%	0.55	0.66	0.47					
LIRMMA	10.66%	9.16%	0.49	0.38	0.68	13.65%	13.04%	0.54	0.52	0.56
LIRMMB	11.81%	9.20%	0.36	0.24	0.73	14.45%	12.62%	0.52	0.41	0.71
LIRMMC	11.78%	9.02%	0.35	0.23	0.71	16.06%	15.02%	0.42	0.28	0.78
LIRMMD	11.32%	8.08%	0.32	0.22	0.57	17.14%	14.31%	0.34	0.22	0.76
LIRMME	10.71%	8.38%	0.37	0.29	0.52	14.89%	12.69%	0.41	0.32	0.59
PEIMEXA	10.30%	7.22%	0.38	0.28	0.62	12.70%	9.25%	0.46	0.39	0.56
PEIMEXB	10.30%	7.61%	0.45	0.37	0.57	12.41%	7.79%	0.64	0.57	0.73
PEIMEXC	10.07%	7.35%	0.37	0.29	0.51	13.42%	10.50%	0.43	0.37	0.51
PEIMEXD	10.11%	7.70%	0.39	0.35	0.44	12.94%	9.86%	0.67	0.61	0.73
PEIMEXE	10.77%	7.32%	0.35	0.25	0.57	12.84%	10.82%	0.31	0.28	0.34
RKMVERIA	10.14%	8.68%	0.52	0.49	0.54	12.17%	8.63%	0.67	0.82	0.56
RKMVERIB	10.66%	9.07%	0.47	0.37	0.65	12.93%	12.31%	0.46	0.81	0.32
RKMVERIC	9.81%	9.08%	0.48	**0.67**	0.38	12.85%	12.85%	0.25	0.86	0.15
RKMVERID	9.97%	8.63%	0.58	0.60	0.56	12.89%	12.89%	0.31	0.80	0.20
RKMVERIE	9.89%	9.28%	0.21	0.35	0.15	12.93%	12.31%	0.46	0.81	0.32
UDCA	10.93%	8.27%	0.26	0.17	0.53					
UDCB	15.79%	11.95%	0.18	0.10	**0.95**					
UDCC	9.47%	8.65%	0.18	0.13	0.29					
UDCD	12.38%	8.54%	0.18	0.11	0.61					
UDCE	9.51%	8.70%	0.18	0.13	0.29					
UNSLA	**8.78%**	7.39%	0.38	0.48	0.32	12.48%	12.00%	0.17	0.57	0.10
UNSLB	8.94%	7.24%	0.40	0.35	0.46	**11.40%**	7.82%	0.61	0.75	0.51
UNSLC	8.82%	6.95%	0.43	0.38	0.49	11.61%	7.82%	0.61	0.75	0.51
UNSLD	10.68%	7.84%	0.45	0.31	0.85	12.93%	9.85%	0.79	**0.91**	0.71
UNSLE	9.86%	7.60%	0.60	0.53	0.70	12.93%	10.13%	0.74	0.90	0.63
UPFA	10.01%	8.28%	0.55	0.56	0.54	13.18%	11.34%	0.72	0.74	0.71
UPFB	10.71%	8.60%	0.48	0.37	0.70	13.01%	11.76%	0.65	0.81	0.54
UPFC	10.26%	9.16%	0.53	0.48	0.61	13.17%	11.60%	0.73	0.76	0.71
UPFD	10.16%	9.79%	0.42	0.42	0.42	12.93%	12.30%	0.60	0.86	0.46
UQAMA	10.04%	7.85%	0.42	0.32	0.62					
TBSA	10.81%	9.22%	0.37	0.29	0.52	13.65%	11.14%	0.67	0.60	0.76
TUA1A	10.19%	9.70%	0.29	0.31	0.27	-	-	0.00	0.00	0.00
TUA1B	10.40%	9.54%	0.27	0.25	0.28	19.90%	19.27%	0.25	0.15	0.76
TUA1C	10.86%	9.51%	0.47	0.35	0.71	13.53%	12.57%	0.36	0.42	0.32
TUA1D	-	-	0.00	0.00	0.00					

Table 2. Performance results achieved by the eRisk 2018 participants (depression and anorexia tasks).

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.

This research has received financial support from the Galician Ministry of Education (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. Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in Twitter. In *ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.
2. David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In *Proceedings Conference and Labs of the Evaluation Forum CLEF 2016*, Evora, Portugal, 2016.
3. David E. Losada, Fabio Crestani, and Javier Parapar. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 346–360, Cham, 2017. Springer International Publishing.
4. David E. Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: Early risk prediction on the internet. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 343–361, Cham, 2018. Springer International Publishing.
5. Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 495–503, New York, NY, USA, 2018. ACM.
6. Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *CoRR*, abs/1804.07000, 2018.