# Overview of eRisk: Early Risk Prediction on the Internet

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, 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, University of A Coruña, A Coruña, Spain
javierparapar@udc.es

**Abstract.** This paper provides an overview of eRisk 2018. This was the second year that this lab was organized at CLEF. The main purpose of eRisk was to explore issues of evaluation methodology, effectiveness metrics and other processes related to early risk detection. Early detection technologies can be employed in different areas, particularly those related to health and safety. The second edition of eRisk had two tasks: a task on early risk detection of depression and a task on early risk detection of anorexia.

## 1 Introduction

The main purpose of this lab is to explore issues of evaluation methodologies, performance metrics and other aspects related to building test collections and defining challenges for early risk detection. Early detection technologies are potentially useful in different areas, particularly those related to safety and health. For example, early alerts could be sent when a person starts showing signs of a mental disorder, when a sexual predator starts interacting with a child, or when a potential offender starts publishing antisocial threats on the Internet. In 2017, our main goal was to pioneer a new interdisciplinary research area that would be potentially applicable to a wide variety of profiles, such as potential paedophiles, stalkers, individuals with a latent tendency to fall into the hands of criminal organisations, people with suicidal inclinations, or people susceptible to depression.

The 2017 lab had two possible ways to participate. One of them followed a classical workshop pattern. This workshop was open to the submission of papers describing test collections or data sets suitable for early risk prediction or early risk prediction challenges, tasks and evaluation metrics. This open submission format was discontinued in 2018. eRisk 2017 also included an exploratory task on early detection of depression. This pilot task was based on the evaluation methodology and test collection presented in a CLEF 2016 paper [1]. The interaction between depression and language use is interesting for early risk detection


P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 343–361, 2018.
https://doi.org/10.1007/978-3-319-98932-7_30

algorithms. We shared this collection with all participating teams and the 2017 participants approached the problem with multiple technologies and models (e.g. Natural Language Processing, Machine Learning, Information Retrieval, etc.). However, the effectiveness of all participating systems was relatively low [2]. For example, the highest F1 was 64%. This suggests that the 2017 task was challenging and there was still much room from improvement.

In 2018, the lab followed a standard campaign-style format. It was composed of two different tasks: early risk detection of depression and early risk detection of anorexia. The first task is a continuation of the eRisk 2017 pilot task. The teams had access to the eRisk 2017 data as training data, and new depression and non-depression test cases were extracted and provided to the participants during the test stage. The second task followed the same format as the depression task. The organizers of the task collected data on anorexia and language use, the data were divided into a training subset and a test subset, and the task followed the same iterative evaluation schedule implemented in 2017 (see below).

## 2 Task 1: Early Detection of Signs of Depression

This is an exploratory task on early detection of signs of depression. The challenge consists of sequentially processing pieces of evidence –in the form of writings posted by depressed or non-depressed users– and learn to detect early signs of depression as soon as possible. The lab focuses on Text Mining solutions and, thus, it concentrates on Social Media submissions (posts or comments in a Social Media website). Texts should be processed by the participating systems in the order they were created. In this way, systems that effectively perform this task could be applied to sequentially track user interactions in blogs, social networks, or other types of online media.

The test collection for this task has the same format as the collection described in [1]. It is a collection of submissions or writings (posts or comments) done by Social Media users. There are two classes of users, depressed and non-depressed. The positive group was obtained by searching for explicit expressions related to a diagnosis (e.g. "diagnosed with depression") and doing a manual check of the retrieved posts. The control group was obtained by random sampling from the large set of social media users available. To make the collection realistic, we also included in the control group users who often post about depression (e.g. individuals who actively participate in the depression threads because they have a close relative suffering from depression). For every user, we collected all his submissions (up to 1000 posts + 1000 comments, which is the limit imposed by the platform), organized them in chronological order, and split this sequence in 10 chunks. The first chunk has the oldest 10% of the submissions, the second chunk has the second oldest 10%, and so forth.

The task was organized into two different stages:

– **Training stage.** Initially, the teams that participated in this task had access to some training data. In this stage, the organizers of the task released the

**Table 1.** Task1 (depression). Main statistics of the train and test collections

| | Train | | Test | |
|---|---|---|---|---|
| | *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 |

entire history of submissions done by a set of training users. All chunks of all training users were sent to the participants. Additionally, the actual class (depressed or non-depressed) of each training user was also provided (i.e. whether or not the user explicitly mentioned that they were diagnosed with depression). In 2018, the training data consisted of all 2017 users (2017 training split + 2017 test split). The participants could therefore tune their systems with the training data and build up from 2017's results. The training dataset was released on Nov 30th, 2017.

– **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. depressed or non-depressed), 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). The first release of test data was done on Feb 6th, 2018 and the last (10th) release of test data was done on April 10th, 2018.

Table 1 reports the main statistics of the train and test collections. The two splits are unbalanced (there are more non-depression cases than depression cases). In the training collection the percentage of depressed cases was about 15% and in the test collection this percentage was about 9%. The number of users is not large, but each user has a long history of submissions (on average, the collections have several hundred submissions per user). Additionally, the mean range of dates from the first submission to the last submission is wide (more than 500 days). Such wide history permits to analyze the evolution of the language from the oldest post or comment to the most recent one.

### 2.1 Evaluation Measures

The evaluation of the tasks considered standard classification measures, such as F1, Precision and Recall (computed with respect to the positive class –depression

or anorexia, respectively–) and an early risk detection measure proposed in [1]. The standard classification measures can be employed to assess the teams' estimations with respect to golden truth judgments that inform us about users that are really positive cases. We include them in our evaluation report because these metrics are well-known and easily interpretable.

However, standard classification measures are time-unaware and do not penalize late decisions. Therefore, the evaluation of the tasks also considered a newer measure of performance that rewards early alerts. More specifically, we employed ERDE, an error measure for early risk detection [1] for which the fewer writings required to make the alert, the better. For each user the evaluation proceeds as follows. Given a chunk of data, if a team's system does not emit a decision then it has access to the next chunk of data (i.e. more submissions from the same user). However, the team's system gets a penalty for *late emission*.

ERDE, which stands for *early risk detection error*, takes into account the correctness of the (binary) decision and the delay taken by the system to make the decision. The delay 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 250 posts or comments (i.e. exactly 25 submissions per chunk to simplify the example). If a team's system emitted a decision for user $u$ after the second chunk of data then the delay $k$ would be 50 (because the system needed to see 50 pieces of evidence in order to make its decision).

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. We will often deal with detection tasks where the number of negative cases is several orders of magnitude larger than the number of positive cases. Hence, if we want to avoid building trivial systems that always say no, we need to have $c_{fn} >> c_{fp}$. 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).

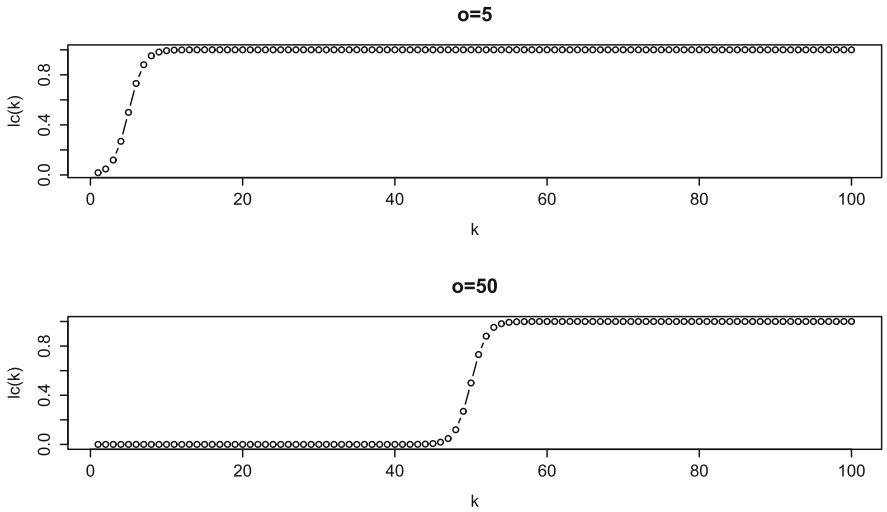


**Fig. 1.** Latency cost functions: $lc_5(k)$ and $lc_{50}(k)$

The function $lc_o(k)$ is a monotonically increasing function of $k$:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{1}$$

The function is parameterised by $o$, which controls the place in the X axis where the cost grows more quickly (Fig. 1 plots $lc_5(k)$ and $lc_{50}(k)$).

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).

To further understand the effect of this penalty let us consider a positive case. Imagine that this positive user is detected by system A after analyzing two texts ($k = 2$), while system B also detects the case but it makes the alert after analyzing 8 texts ($k = 8$). $ERDE_5$ would assign system A an error of 0.047 and system B an error of 0.9526.

All cost weights are in $[0, 1]$ and, thus, ERDE is in the range $[0, 1]$. Systems had to take one decision for each subject and the overall error is the mean of the $p$ ERDE values.

### 2.2 Results

Each team could submit up to 5 runs or variants. We received 45 contributions from 11 different institutions. This is a substantial increase with respect to erisk

2017, which had 8 institutions and 30 contributed runs. Table 3 reports the institutions that contributed to eRisk 2018 and the labels associated to their runs.

First, let us analyze the behaviour of the systems in terms of how fast they emitted decisions. Figure 2 shows a boxplot graph of the number of chunks required to make the decisions. The test collection has 820 users and, thus, each boxplot represents the statistics of 820 cases.

Some systems (RKMVERIB, RKMVERIC, RKMVERID, RKMVERIE, TBSA, UPFC, UPFD) took all decisions after the last chunk (i.e. did not emit any earlier decision). These variants were extremely conservative: they waited to see the whole history of submissions for all users and, next, they emitted their decisions. Remember that all teams were forced to emit a decision for each user at the last chunk.

Many other runs also took most of the decisions after the last chunk. For example, FHDO-BCSGA assigned a decision at the last chunk in 725 out of 820 users. Only a few runs were really quick at emitting decisions. Notably, most UDC's runs and LIIRA had a median of 1 chunk needed to emit decisions.

Figure 3 shows a boxplot of the number of submissions required by each run in order to emit decisions. Most of the time the teams waited to see hundreds of writings for each user. Only a few submissions (UDCA, UDCB, UDCD, UDCE, UNSLD, some LIIRx runs) had a median number of writings analyzed below 100. It appears that the teams have concentrated on accuracy (rather than delay) and, thus, they did not care much about penalties for late decisions. A similar behaviour was found in the runs submitted in 2017.

The number of user submissions has a high variance. Some users have only 10 submissions, while other users have more than a thousand submissions. It would be interesting to study the interaction between the number of user submissions and the effectiveness of the estimations done by the participating systems. This study could help to shed light on issues such as the usefulness of a large (vs short) history of submissions and the effect of off-topic submissions (e.g. submissions totally unrelated to depression).

Another intriguing issue relates to potential false positives. For instance, a doctor who is active on the depression community because he gives support to people suffering from depression, or a wife whose husband has been diagnosed with depression. These people would often write about depression and possibly use a style that might imply they are depressed, but obviously they are not. The collection contains this type of non-depressed users and these cases are challenging for automatic classification. Arguably, these non-depressed users are much different from other non-depressed users who do not engage in any depression-related conversation. In any case, this issue requires further investigation. For example, it will be interesting to do error analysis with the systems' decisions and check the characteristics of the false positives.

Figure 4 helps to analyze another aspect of the decisions emitted by the teams. For each user class, it plots the percentage of correct decisions against the number of users. For example, the last two bars of the upper plot show that about 5 users were correctly identified by more than 90% of the runs. Similarly,

the rightmost bar of the lower plot means that a few non-depressed users were correctly classified by all runs (100% correct decisions). The graphs show that the teams tend to be more effective with non-depressed users. This is as expected because most non-depressed cases do not engage in depression-related conversations and, therefore, they are easier to distinguish from depressed users. The distribution of correct decisions for non-depressed users has many cases where more than 80% of the systems are correct. The distribution of correct decisions for depressed users is flatter, and many depressed users are only identified by a low percentage of the runs. This suggests that the teams implemented a wide range of strategies that detect different portions of the depression class. Furthermore, there are not depressed users that are correctly identified by all systems. However, an interesting point is that no depressed user has 0% of correct decisions. This means that every depressed user was classified as such by at least one run. In the future, it will be interesting to perform error analysis and try to understand why some positive cases are really hard to detect (e.g. is it that we have little evidence on such cases?).

Let us now analyze the effectiveness results (see Table 4). The first conclusion we can draw is that the task is as difficult as in 2017. In terms of F1, performance is again low. The highest F1 is 0.64 and the highest precision is 0.67. This might be related to the effect of false positives discussed above. The lowest $ERDE_{50}$ was achieved by the FHDO-BCSG team, which also submitted the runs that performed the best in terms of $F1$. The run with the lowest $ERDE_5$ was submitted by the UNSLA team and the run with the highest precision was submitted by RKMVERI. The UDC team submitted a high recall run (0.95) but its precision was extremely low.

In terms of $ERDE_5$, the best performing run is UNSLA, which has poor F1, Precision and Recall. This run was not good at identifying many depressed users but, still, it has low $ERDE_5$. This suggests that the true positives were emitted by this run at earlier chunks (quick emissions). $ERDE_5$ is extremely stringent with delays (after 5 writings, penalties grow quickly, see Fig. 1). This promotes runs that emit few but quick depression decisions. $ERDE_{50}$, instead, gives smoother penalties to delays. This makes that the run with the lowest $ERDE_{50}$, FHDO-BCSGB, has much higher F1 and Precision. Such difference between $ERDE_5$ and $ERDE_{50}$ is highly relevant in practice. For example, a mental health agency seeking an automatic tool for screening depression could set the penalty weights depending on the consequences of late detection of signs of depression.

## 3 Task 2: Early Detection of Signs of Anorexia

Task 2 was an exploratory task on early detection of signs of anorexia. The format of the task, data extraction methods and evaluation methodology (training stage followed by a test stage with on sequential releases of user data) was the same used for Task 1. This task was introduced in 2018 and, therefore, all users (training + test) were collected just for this new task.

**Table 2.** Task2 (anorexia). Main statistics of the train and test collections

| | Train | | Test | |
|---|---|---|---|---|
| | *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 2 reports the main statistics of the train and test collections of Task 2. The collection shares the main characteristics of Task 1's collections: the two splits are unbalanced (of course, there are more non-anorexia cases than anorexia cases). Contrary to the depression case, the number of users is not large (and, again, each user has a long history of submissions). The mean range of dates from the first submission to the last submission is also wide (more than 500 days).

### 3.1 Results

Each team could submit up to 5 runs or variants. We received 35 contributions from 9 different institutions. All institutions participating in Task 2 had also sent results for Task 1. Table 5 reports the institutions that contributed to this second task of eRisk 2018 and the labels associated to their runs.

The behaviour of the systems in terms of how fast they emitted decisions is shown in Fig. 5, which includes boxplot graphs of the number of chunks required to make the decisions. The test collection of Task 2 has 320 users and, thus, each boxplot represents the statistics of 320 cases. The trends are similar to those found in Task 1. Mosf of the systems emitted decisions at a late stage with only a few exceptions (notably, LIIRA and LIIRB). LIIRA and LIIRB had a median number of chunks analyzed of 3 and 6, respectively. The rest of the systems had a median number of chunks analized equal to or near 10.

Figure 6 shows a boxplot of the number of submissions required by each run in order to emit decisions. Again, most of the variants analyzed hundred of submissions before emitting decisions. Only the two LIIR runs discussed above and LIRMMD opted for emitting decisions after a fewer number of user submissions. In Task 2, again, most of the teams have ignored the penalties for late decisions and they have mostly focused on classification accuracy.

Figure 7 plots the percentage of correct decisions against the number of users. The plot shows again a clear distinction between the positive class (anorexia) and the negative class (non-anorexia). Most of the non-anorexia users are correctly identified by most of the systems (nearly all non-anorexia users fall in the range

**Table 3.** Task 1 (depression). Participating institutions and submitted results

| Institution | Submitted files |
|---|---|
| FH Dortmund, Germany | FHDO-BCSGA |
| | FHDO-BCSGB |
| | FHDO-BCSGC |
| | FHDO-BCSGD |
| | FHDO-BCSGE |
| IRIT, France | LIIRA |
| | LIIRB |
| | LIIRC |
| | LIIRD |
| | LIIRE |
| LIRMM, University of Montpellier, France | LIRMMA |
| | LIRMMB |
| | LIRMMC |
| | LIRMMD |
| | LIRMME |
| Instituto Tecnológico Superior del Oriente del Estado de Hidalgo, Mexico<br>Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico<br>Universidad de Houston, USA<br>& Universidad Autónoma del Estado de Hidalgo, Mexico | PEIMEXA |
| | PEIMEXB |
| | PEIMEXC |
| | PEIMEXD |
| | PEIMEXE |
| Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math, West Bengal, India | RKMVERIA |
| | RKMVERIB |
| | RKMVERIC |
| | RKMVERID |
| | RKMVERIE |
| University of A Coruña, Spain | UDCA |
| | UDCB |
| | UDCC |
| | UDCD |
| | UDCE |
| Universidad Nacional de San Luis, Argentina | UNSLA |
| | UNSLB |
| | UNSLC |
| | UNSLD |
| | UNSLE |
| Universitat Pompeu Fabra, Spain | UPFA |
| | UPFB |
| | UPFC |
| | UPFD |
| Université du Québec à Montréal, Canada | UQAMA |
| The Black Swan, Taiwan | TBSA |
| Tokushima University, Japan | TUA1A |
| | TUA1B |
| | TUA1C |
| | TUA1D |

80%–100%, meaning that at least 80% of the systems labeled them as non-anorexic). In contrast, the distribution of anorexia users is flatter and, in many cases, they are only identified by less than half of the systems. An interesting result is that all anorexia users were identified by at least 10% of the systems.

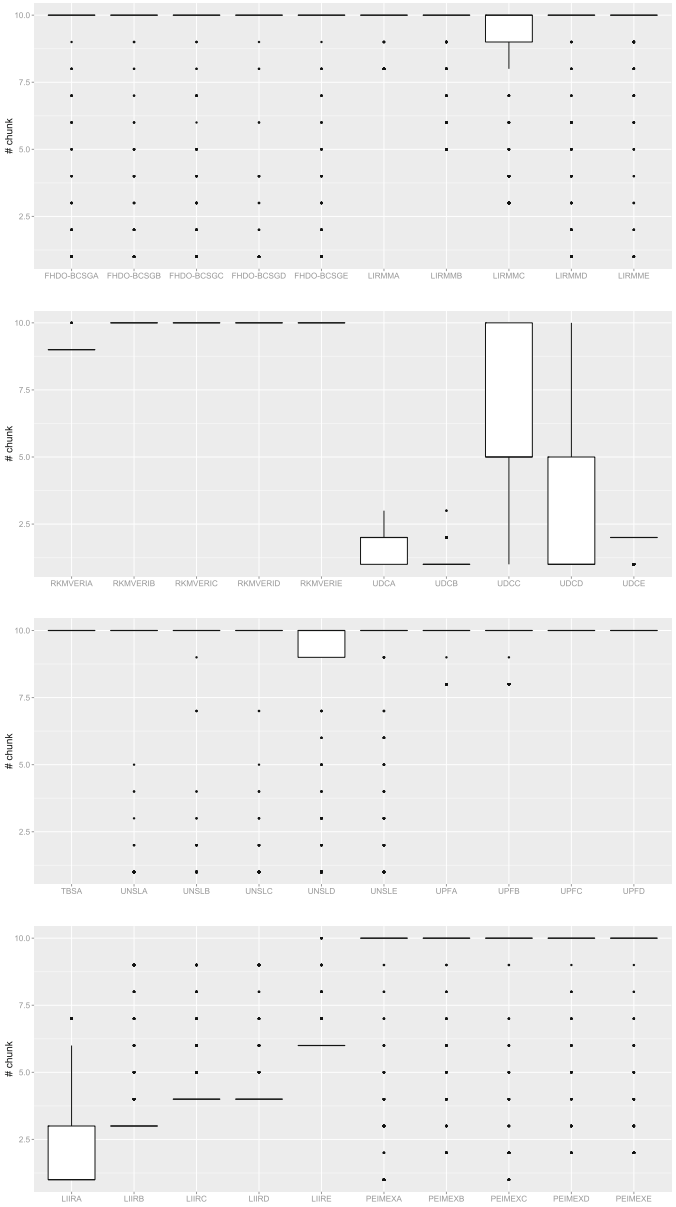


**Fig. 2.** Number of chunks required by each contributing run in order to emit a decision.

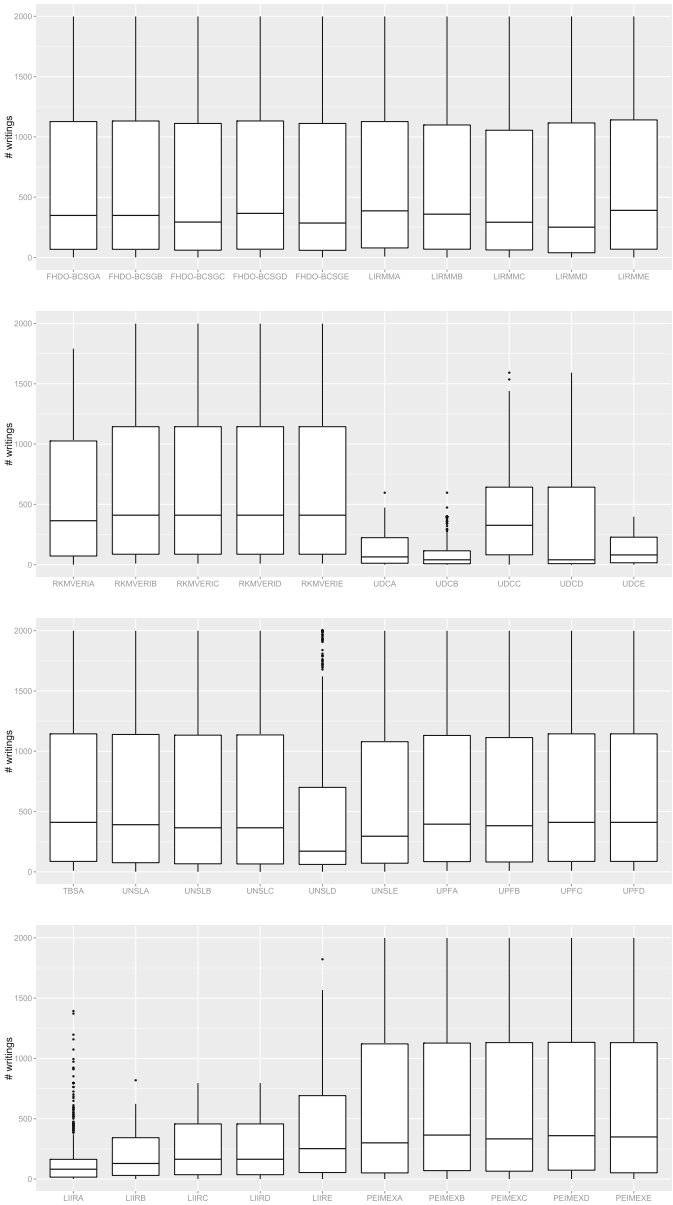


**Fig. 3.** Number of writings required by each contributing run in order to emit a decision.

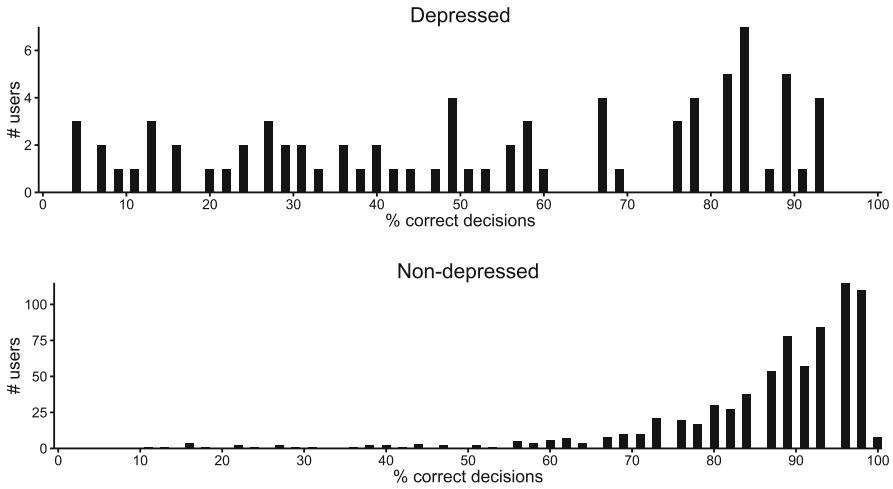


**Fig. 4.** Number of depressed and non-depressed subjects that had a given percentage of correct decisions.

Table 6 reports the effectiveness of the systems. In general, performance is remarkably higher than that achieved by the systems for Task 1. There could be a number of reasons for such an outcome. First, the proportion of potential false positives (e.g. people engaging in anorexia-related conversations) might be lower in Task 2's test collection. This hypothesis would need to be investigated through a careful analysis of the data. Second, the submissions of anorexia users might be extremely focused on eating habits, losing weights, etc. If they do not often engage in general (anorexia unrelated) conversations then it would be easier for the systems to distinguish them from other users. In any case, these are only speculations and this issue requires further research.

The highest F1 is 0.85 and the highest precision is 0.91. The lowest $ERDE_{50}$ was achieved by FHDO-BCSGD, which also has the highest recall (0.88). The run with the lowest $ERDE_5$ was submitted by the UNSL team (UNSLB), which shows again that this team paid more attention to emitting early decisions (at least for the true positives).

Overall, the results obtained by the teams are promising. The high performance achieved suggest that it is feasible to design automatic text analysis tools that make early alerts of signs of eating disorders.

**Table 4.** Task 1 (depression). Results

| | $ERDE_5$ | $ERDE_{50}$ | F1 | P | R |
|---|---|---|---|---|---|
| FHDO-BCSGA | 9.21% | 6.68% | 0.61 | 0.56 | 0.67 |
| FHDO-BCSGB | 9.50% | **6.44%** | **0.64** | 0.64 | 0.65 |
| FHDO-BCSGC | 9.58% | 6.96% | 0.51 | 0.42 | 0.66 |
| FHDO-BCSGD | 9.46% | 7.08% | 0.54 | 0.64 | 0.47 |
| FHDO-BCSGE | 9.52% | 6.49% | 0.53 | 0.42 | 0.72 |
| LIIRA | 9.46% | 7.56% | 0.50 | 0.61 | 0.42 |
| LIIRB | 10.03% | 7.09% | 0.48 | 0.38 | 0.67 |
| 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 |
| LIRMMB | 11.81% | 9.20% | 0.36 | 0.24 | 0.73 |
| LIRMMC | 11.78% | 9.02% | 0.35 | 0.23 | 0.71 |
| LIRMMD | 11.32% | 8.08% | 0.32 | 0.22 | 0.57 |
| LIRMME | 10.71% | 8.38% | 0.37 | 0.29 | 0.52 |
| PEIMEXA | 10.30% | 7.22% | 0.38 | 0.28 | 0.62 |
| PEIMEXB | 10.30% | 7.61% | 0.45 | 0.37 | 0.57 |
| PEIMEXC | 10.07% | 7.35% | 0.37 | 0.29 | 0.51 |
| PEIMEXD | 10.11% | 7.70% | 0.39 | 0.35 | 0.44 |
| PEIMEXE | 10.77% | 7.32% | 0.35 | 0.25 | 0.57 |
| RKMVERIA | 10.14% | 8.68% | 0.52 | 0.49 | 0.54 |
| RKMVERIB | 10.66% | 9.07% | 0.47 | 0.37 | 0.65 |
| RKMVERIC | 9.81% | 9.08% | 0.48 | **0.67** | 0.38 |
| RKMVERID | 9.97% | 8.63% | 0.58 | 0.60 | 0.56 |
| RKMVERIE | 9.89% | 9.28% | 0.21 | 0.35 | 0.15 |
| 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 |
| UNSLB | 8.94% | 7.24% | 0.40 | 0.35 | 0.46 |
| UNSLC | 8.82% | 6.95% | 0.43 | 0.38 | 0.49 |
| UNSLD | 10.68% | 7.84% | 0.45 | 0.31 | 0.85 |
| UNSLE | 9.86% | 7.60% | 0.60 | 0.53 | 0.70 |
| UPFA | 10.01% | 8.28% | 0.55 | 0.56 | 0.54 |
| UPFB | 10.71% | 8.60% | 0.48 | 0.37 | 0.70 |
| UPFC | 10.26% | 9.16% | 0.53 | 0.48 | 0.61 |
| UPFD | 10.16% | 9.79% | 0.42 | 0.42 | 0.42 |
| UQAMA | 10.04% | 7.85% | 0.42 | 0.32 | 0.62 |
| TBSA | 10.81% | 9.22% | 0.37 | 0.29 | 0.52 |
| TUA1A | 10.19% | 9.70% | 0.29 | 0.31 | 0.27 |
| TUA1B | 10.40% | 9.54% | 0.27 | 0.25 | 0.28 |
| TUA1C | 10.86% | 9.51% | 0.47 | 0.35 | 0.71 |
| TUA1D | - | - | 0.00 | 0.00 | 0.00 |

**Table 5.** Task 2 (anorexia). Participating institutions and submitted results

| Institution | Submitted files |
|---|---|
| FH Dortmund, Germany | FHDO-BCSGA |
| | FHDO-BCSGB |
| | FHDO-BCSGC |
| | FHDO-BCSGD |
| | FHDO-BCSGE |
| IRIT, France | LIIRA |
| | LIIRB |
| LIRMM, University of Montpellier, France | LIRMMA |
| | LIRMMB |
| | LIRMMC |
| | LIRMMD |
| | LIRMME |
| Instituto Tecnológico Superior del Oriente del Estado de Hidalgo, Mexico | PEIMEXA |
| Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico | PEIMEXB |
| Universidad de Houston, USA & Universidad Autónoma del Estado de Hidalgo, Mexico | PEIMEXC |
| | PEIMEXD |
| | PEIMEXE |
| Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math, West Bengal, India | RKMVERIA |
| | RKMVERIB |
| | RKMVERIC |
| | RKMVERID |
| | RKMVERIE |
| Universidad Nacional de San Luis, Argentina | UNSLA |
| | UNSLB |
| | UNSLC |
| | UNSLD |
| | UNSLE |
| Universitat Pompeu Fabra, Spain | UPFA |
| | UPFB |
| | UPFC |
| | UPFD |
| The Black Swan, Taiwan | TBSA |
| Tokushima University, Japan | TUA1A |
| | TUA1B |
| | TUA1C |

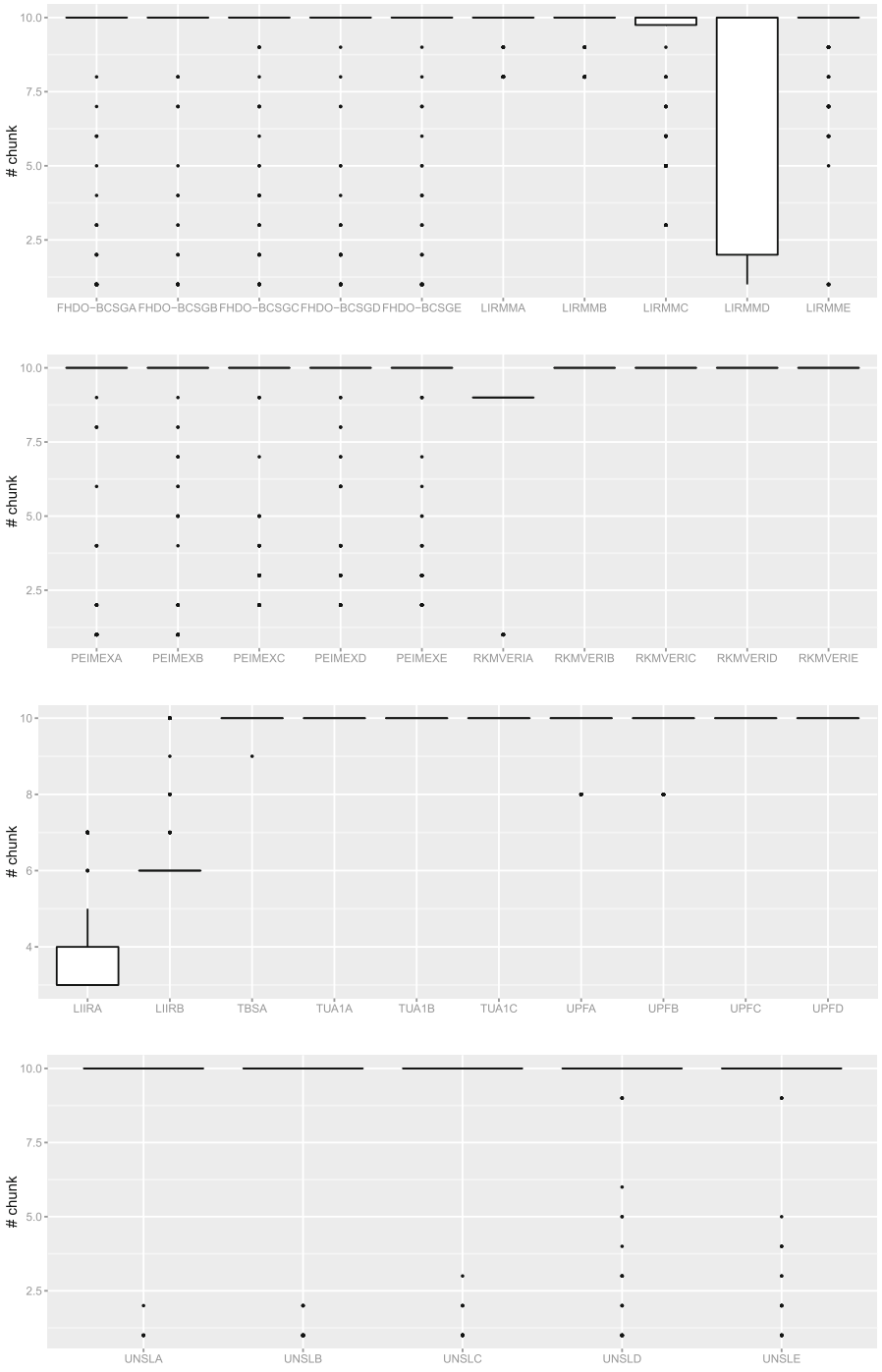


**Fig. 5.** Number of chunks required by each contributing run in order to emit a decision.

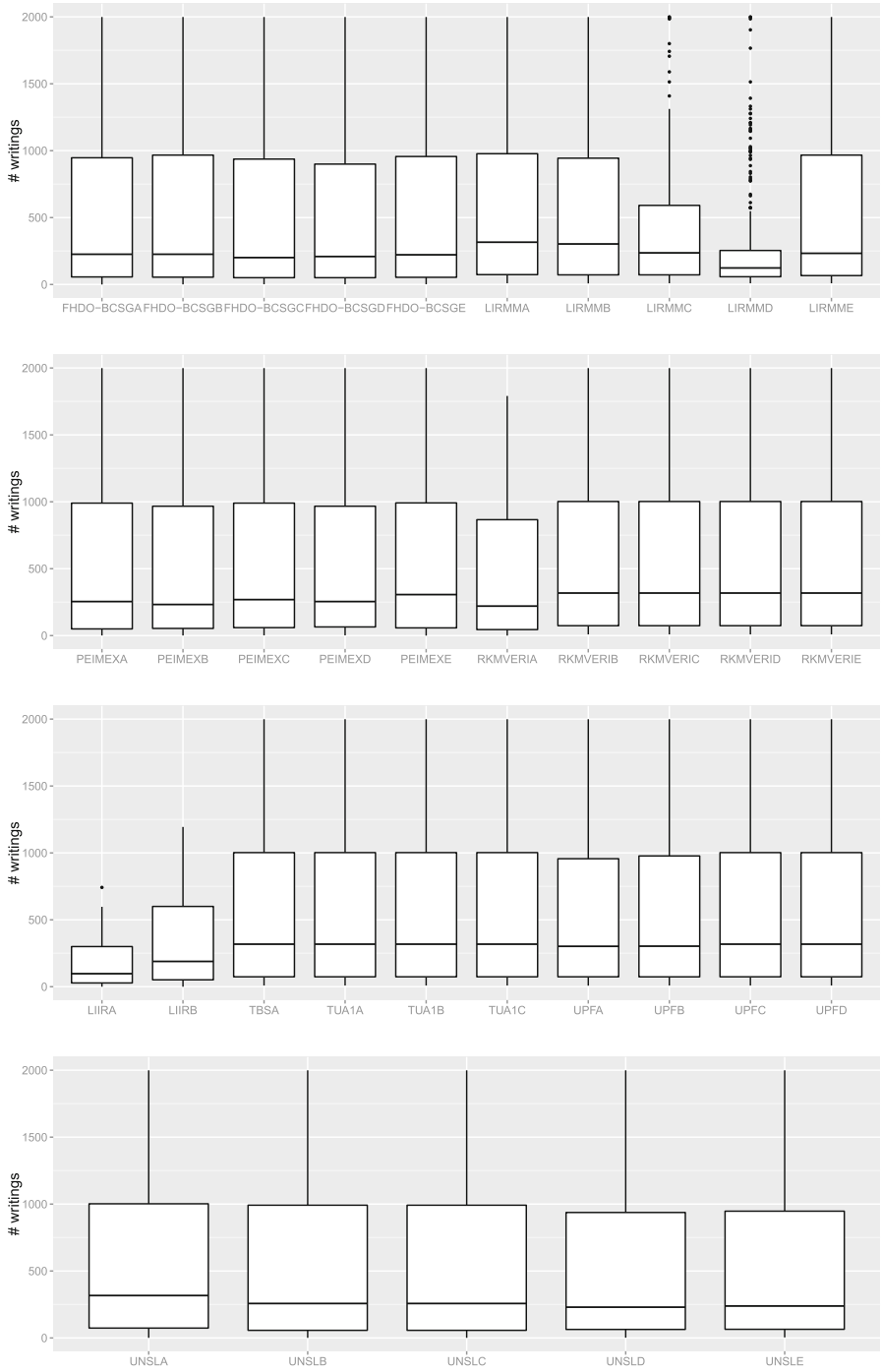


**Fig. 6.** Number of writings required by each contributing run in order to emit a decision.

**Table 6.** Task 2 (anorexia). Results

| | $ERDE_5$ | $ERDE_{50}$ | F1 | P | R |
|---|---|---|---|---|---|
| FHDO-BCSGA | 12.17% | 7.98% | 0.71 | 0.67 | 0.76 |
| FHDO-BCSGB | 11.75% | 6.84% | 0.81 | 0.84 | 0.78 |
| FHDO-BCSGC | 13.63% | 9.64% | 0.55 | 0.47 | 0.66 |
| FHDO-BCSGD | 12.15% | **5.96%** | 0.81 | 0.75 | **0.88** |
| FHDO-BCSGE | 11.98% | 6.61% | **0.85** | 0.87 | 0.83 |
| LIIRA | 12.78% | 10.47% | 0.71 | 0.81 | 0.63 |
| LIIRB | 13.05% | 10.33% | 0.76 | 0.79 | 0.73 |
| LIRMMA | 13.65% | 13.04% | 0.54 | 0.52 | 0.56 |
| LIRMMB | 14.45% | 12.62% | 0.52 | 0.41 | 0.71 |
| LIRMMC | 16.06% | 15.02% | 0.42 | 0.28 | 0.78 |
| LIRMMD | 17.14% | 14.31% | 0.34 | 0.22 | 0.76 |
| LIRMME | 14.89% | 12.69% | 0.41 | 0.32 | 0.59 |
| PEIMEXA | 12.70% | 9.25% | 0.46 | 0.39 | 0.56 |
| PEIMEXB | 12.41% | 7.79% | 0.64 | 0.57 | 0.73 |
| PEIMEXC | 13.42% | 10.50% | 0.43 | 0.37 | 0.51 |
| PEIMEXD | 12.94% | 9.86% | 0.67 | 0.61 | 0.73 |
| PEIMEXE | 12.84% | 10.82% | 0.31 | 0.28 | 0.34 |
| RKMVERIA | 12.17% | 8.63% | 0.67 | 0.82 | 0.56 |
| RKMVERIB | 12.93% | 12.31% | 0.46 | 0.81 | 0.32 |
| RKMVERIC | 12.85% | 12.85% | 0.25 | 0.86 | 0.15 |
| RKMVERID | 12.89% | 12.89% | 0.31 | 0.80 | 0.20 |
| RKMVERIE | 12.93% | 12.31% | 0.46 | 0.81 | 0.32 |
| UNSLA | 12.48% | 12.00% | 0.17 | 0.57 | 0.10 |
| UNSLB | **11.40%** | 7.82% | 0.61 | 0.75 | 0.51 |
| UNSLC | 11.61% | 7.82% | 0.61 | 0.75 | 0.51 |
| UNSLD | 12.93% | 9.85% | 0.79 | **0.91** | 0.71 |
| UNSLE | 12.93% | 10.13% | 0.74 | 0.90 | 0.63 |
| UPFA | 13.18% | 11.34% | 0.72 | 0.74 | 0.71 |
| UPFB | 13.01% | 11.76% | 0.65 | 0.81 | 0.54 |
| UPFC | 13.17% | 11.60% | 0.73 | 0.76 | 0.71 |
| UPFD | 12.93% | 12.30% | 0.60 | 0.86 | 0.46 |
| TBSA | 13.65% | 11.14% | 0.67 | 0.60 | 0.76 |
| TUA1A | - | - | 0.00 | 0.00 | 0.00 |
| TUA1B | 19.90% | 19.27% | 0.25 | 0.15 | 0.76 |
| TUA1C | 13.53% | 12.57% | 0.36 | 0.42 | 0.32 |

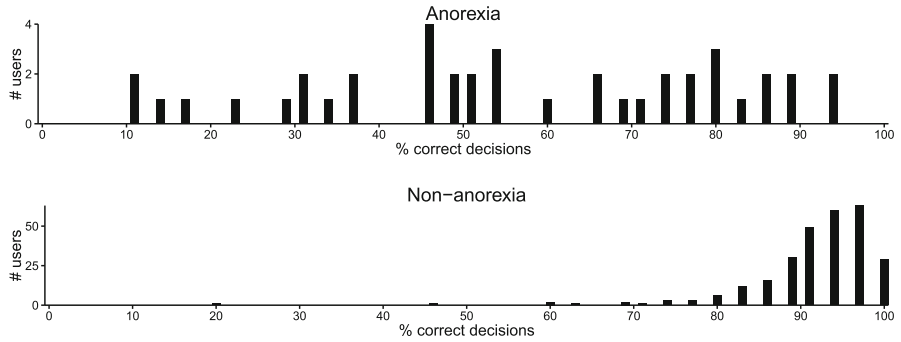


**Fig. 7.** Number of anorexia and non-anorexia users that had a given percentage of correct decisions.

## 4 Conclusions

This paper provided an overview of eRisk 2018. This was the second year that this lab was organized at CLEF and the lab's activities concentrated on two tasks (early detection of signs of depression and early detection of signs of anorexia). Overall, the tasks received 80 variants or runs and the teams focused on tuning different classification solutions. The tradeoff between early detection and accuracy was ignored by most participants.

The effectiveness of the solutions implemented to early detect signs of depression is similar to that achieved for eRisk 2017. This performance is still modest, suggesting that it is challenging to tell depressed and non-depressed users apart. In contrast, the effectiveness of the systems that detect signs of anorexia was much higher. This promising result encourages us to further explore the creation of benchmarks for text-based screening of eating disorders. In the future, we also want to instigate more research on the tradeoff between accuracy and delay.

**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 Economía y Competitividad" of the Government of Spain and FEDER Funds under the research project TIN2015-64282-R, (ii) Xunta de Galicia (project GPC 2016/035), and (iii) Xunta de Galicia – "Consellería de Cultura, Educación e Ordenación Universitaria" and the European Regional Development Fund (ERDF) through the following 2016–2019 accreditations: ED431G/01 ("Centro singular de investigacion de Galicia") and ED431G/08.

## References

1. 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
2. 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