Overview of eRisk at CLEF 2019 Early Risk Prediction on the Internet (extended overview)

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, Centro de Investigación en Tecnologías de la Información y las Comunicaciones, Universidade da Coruña, Spain
javierparapar@udc.es

Abstract. This paper provides an overview of eRisk 2019, the third edition of this lab under the CLEF conference. The main purpose of eRisk is 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. This edition of eRisk had three tasks. Two of them shared the same format and focused on early detecting signs of depression (T1) or self-harm (T2). The third task (T3) focused on an innovative challenge related to automatically filling a depression questionnaire based on user interactions in social media.

1 Introduction

The main purpose of eRisk 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.

Although the evaluation methodology (strategies to build new test collections, novel evaluation metrics, etc) can be applied on multiple domains, eRisk has so far focused on psychological problems (essentially, depression, self-harm and eating disorders). In 2017 [4, 5], we ran an exploratory task on early detection of depression. This pilot task

was based on the evaluation methodology and test collection presented in [3]. In 2018 [7, 6], we ran a continuation of the task on early detection of signs of depression together with a new task on early detection of signs of anorexia. Over these years, we have been able to compare a number of solutions that employ multiple technologies and models (e.g. Natural Language Processing, Machine Learning, or Information Retrieval). We learned that the interaction between psychological problems and language use is challenging and, in general, the effectiveness of most contributing systems is modest. For example, in terms of detecting signs of depression, the highest F1 was about 65%. This suggests that this kind of early prediction tasks require further research and the solutions proposed so far still have much room from improvement.

In 2019, the lab had three campaign-style tasks. Two of them had the same orientation of previous eRisk tasks but we changed the way in which data was released and, additionally, we expanded the set of evaluation measures. These two tasks were oriented to early detection of signs of anorexia and self-harm, respectively. The third task, which was completely new, was oriented to analyzing a user's history of posts and extracting useful evidence for estimating the user's depression level. More specifically, the participants had to process the user's posts and, next, estimate the user's answers to a standard depression questionnaire. These three tasks are described in the next sections of this overview paper.

2 Task 1: Early Detection of Signs of Anorexia

This is the continuation of eRisk 2018's T2 task. The challenge consists of sequentially processing pieces of evidence and detect early traces of anorexia as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media. Texts had to be processed in the order they were posted. In this way, systems that effectively perform this task could be applied to sequentially monitor user interactions in blogs, social networks, or other types of online media.

The test collection was built using the same methodology and sources as the collection described in [3]. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, those affected or not by anorexia, and, for each user, the collection contains a sequence of writings (in chronological order). The positive set is composed of users who explicitly mentioned that they were diagnosed with anorexia[1], while the negative set is mainly composed of random users from the same social media platform. To make the collection realistic, we also included in the negative group users who often post about anorexia (e.g. individuals who actively participate in the anorexia threads because they have a close relative suffering from this eating disorder). For every user, we collected all his submissions (up to 1000 posts and 1000 comments, which is the limit imposed by the platform), and organized them in chronological order.

The task was organized into two different stages:

[1] However, following the extraction method suggested by Coppersmith and colleagues[2], the post discussing the diagnosis was removed from the collection.

Table 1. T1 (anorexia). Main statistics of the collection

	Train		Test	
	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

- **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 entire history of submissions done by a set of training users. In 2019, the training data consisted of 2018's T2 data (2018 training split + 2018 test split). The participants could therefore tune their systems with the training data and build up from 2018's results. The training dataset was released on Nov 30th, 2018.
- **Test stage**. In 2019, we moved from a "chunk-based" release of test data (used in 2017 and 2018) to a "item-by-item" release of test data. We set up a server that iteratively gave user writings to the participating teams[2]. In this way, each participant had the opportunity to stop and make an alert at any point of the user chronology. After reading each user post, the teams had to choose between: i) emitting an alert on the user, or ii) making no alert on the user. Alerts were considered as final (i.e. further decisions about this individual were ignored), while *no alerts* were considered as non-final (i.e. the participants could later submit an alert for this user if they detected the appearance of risk signs). This choice had to be made for each user in the test split. The systems were evaluated based on the accuracy of the decisions and the number of user writings required to take the decisions (see below). A REST server was built to support the test stage. The server iteratively gave user writings to the participants and waited for their responses (no new user data provided until the system said alert/no alert). This server was running from March 3rd, 2019 to April 10th, 2019.

Table 1 reports the main statistics of the train and test collections used for T1. In 2019, we also decided to expand the toolkit of evaluation measures. This is discussed next.

2.1 Decision-based Evaluation

This form of evaluation revolves around the (binary) decisions taken for each user by the participating systems. Besides standard classification measures (Precision, Recall and F1[3]), we computed $ERDE$, the early risk detection error used in the previous editions of the lab. A full description of $ERDE$ can be found in [3]. Essentially, $ERDE$ is

[2] More information about the server can be found on the lab website http://early.irlab.org/server.html

[3] computed with respect to the positive class.

an error measure that introduces a penalty for late correct alerts (true positives). The penalty grows with the delay in emitting the alert, and the delay is measured here as the number of user posts that had to be processed before making the alert.

In 2019, we complemented the evaluation report with additional decision-based metrics that try to capture additional aspects of the problem. These metrics try to overcome some limitations of $ERDE$, namely:

- the penalty associated to true positives goes quickly to 1. This is due to the functional form of the cost function (sigmoid).
- a perfect system, which detects the true positive case right after the first round of messages (first chunk), does not get error equal to 0.
- with a method based on releasing data in a chunk-based way (as it was done in 2017 and 2018) the contribution of each user to the performance evaluation has a large variance (different for users with few writings per chunk vs users with many writings per chunk).
- $ERDE$ is not interpretable.

Some research teams have analysed these issues and proposed alternative ways for evaluation. Trotzek and colleagues [9] proposed $ERDE_o^{\%}$. This is a variant of ERDE that does not depend on the number of user writings seen before the alert but, instead, it depends on the *percentage* of user writings seen before the alert. In this way, user's contributions to the evaluation are normalized (currently, all users weight the same). However, there is an important limitation of $ERDE_o^{\%}$. In real life applications, the overall number of user writings is not known in advance. Social Media users post contents online and screening tools have to make predictions with the evidence seen. In practice, you do not know when (and if) a user's thread of message is exhausted. Thus, the performance metric should not depend on such lack of knowledge of the total number of user writings.

Another proposal of an alternative evaluation metric for early risk prediction was done by Sadeque and colleagues [8]. They proposed $F_{latency}$, which fits better with our purposes. This measure is described next.

Imagine a user $u \in U$ and an early risk detection system that iteratively analyzes u's writings (e.g. in chronological order, as they appear in Social Media) and, after analyzing k_u user writings ($k_u \geq 1$), takes a binary decision $d_u \in \{0, 1\}$, which represents the decision of the system about the user being a risk case. By $g_u \in \{0, 1\}$, we refer to the user's golden truth label. A key component of an early risk evaluation should be the delay on detecting true positives (we do not want systems to detect these cases too late). Therefore, a first and intuitive measure of delay can be defined as follows[4]:

$$\text{latency}_{TP} = \text{median}\{k_u : u \in U, d_u = g_u = 1\} \quad (1)$$

[4] Observe that Sadeque et al (see [8], pg 497) computed the latency for all users such that $g_u = 1$. We argue that latency should be computed only for the true positives. The false negatives ($g_u = 1$, $d_u = 0$) are not detected by the system and, therefore, they would not generate an alert.

This measure of latency goes over the true positives detected by the system and assesses the system's delay based on the median number of writings that the system had to process to detect such positive cases. This measure can be included in the experimental report together with standard measures such as Precision (P), Recall (R) and the F-measure (F):

$$P = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : d_u = 1|} \tag{2}$$

$$R = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : g_u = 1|} \tag{3}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

Furthermore, Sadeque et al. proposed a measure, $F_{latency}$, which combines the effectiveness of the decision (estimated with the F measure) and the delay[5]. This is based on multiplying F by a penalty factor based on the median delay. More specifically, each individual (true positive) decision, taken after reading k_u writings, is assigned the following penalty:

$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}} \tag{5}$$

where p is a parameter that determines how quickly the penalty should increase. In [8], p was set such that the penalty equals 0.5 at the median number of posts of a user[6]. Observe that a decision right after the first writing has no penalty ($penalty(1) = 0$). Figure 1 plots how the latency penalty increases with the number of observed writings.

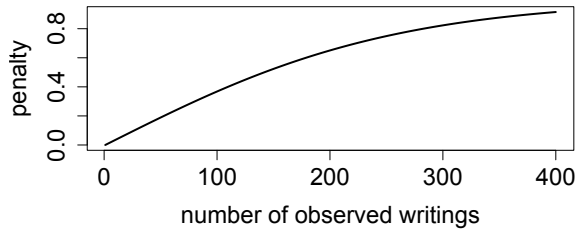

Fig. 1. Latency penalty increases with the number of observed writings (k_u)

The system's overall speed factor is computed as:

[5] Again, we adopt Sadeque et al.'s proposal but we estimate latency only over the true positives.
[6] In the eRisk 2017 collection this led to setting p to 0.0078.

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\}) \quad (6)$$

speed equals 1 for a system whose true positives are detected right at the first writing. A slow system, which detects true positives after hundreds of writings, will be assigned a speed score near 0.

Finally, the *latency-weighted* F score is simply:

$$F_{latency} = F \cdot speed \quad (7)$$

In 2019, user's data was processed by the participants in a post by post basis (i.e. we avoided a chunk-based release of data). Under these conditions, the evaluation approach has the following nice properties:

- smooth grow of penalties.
- a perfect system gets $F_{latency} = 1$.
- for each user u the system can opt to stop at any point k_u and, therefore, now we do not have the effect of an imbalanced importance of users.
- $F_{latency}$ is more interpretable than $ERDE$.

2.2 Ranking-based Evaluation

This section discusses an alternative form of evaluation, which was used as a complement of the evaluation described above. After each release of data (new user writing) the participants had to send back the following information (for each user in the collection): i) a decision for the user (alert/no alert), which was used to compute the decision-based metrics discussed above, and ii) a score that represents the user's level of risk (estimated from the evidence seen so far). We used these scores to build a ranking of users in decreasing estimation of risk. For each participating system, we have one ranking at each point (i.e. ranking after 1 writing, ranking after 2 writings, etc.). This simulates a continuous re-ranking approach based on the evidence seen so far. In a real life application, this ranking would be presented to an expert user who could take decisions (e.g. by inspecting the rankings).

Each ranking can be scored with standard IR metrics, such as P@10 or NDCG. We therefore report the ranking-based performance of the systems after seeing k writings (with varying k).

2.3 Participants

Table 2 reports the participating groups and the runs that they submitted for each eRisk task. The next paragraphs give a brief summary on the techniques implemented by each of them.

UppsalaNLP: This is a joint collaboration between Uppsala University, Gavagai, and the KTH Royal Institute of Technology (Sweden). They evaluated different types of

Table 2. eRisk 2019. Participants

team	T1 #runs	T2 #runs	T3 #runs
UppsalaNLP	5	-	-
BioInfo@UAVR	1	1	1
BiTeM	5	5	5
lirmm	5	5	5
CLaC	5	-	-
SINAI	3	-	-
HULAT	5	-	-
UDE	5	5	5
SSN-NLP	5	-	-
Fazl	3	3	3
UNSL	5	5	5
LTL-INAOE	2	4	4
INAOE-CIMAT	5	-	-
CAMH	-	5	5

semantic vectors or word embeddings for text representation and two different classification architectures (linear regression and multi-layer perceptrons). They represented the lexical items in the posts under analysis as word embeddings and tested three semantic vector models: Random Indexing, which is based on the Sparse Distributed Memory model, GloVe, which is used in a broad range of academic experiments, and the recently published ELMo, which has shown to provide solid representations for general purpose applications.

BioInfo@UAVR: This team, from the Bioinformatics group of the Institute of Electronics and Engineering Informatics of University of Aveiro (Portugal), designed a mixed approach that combines machine learning with psycholinguistics and behavioural patterns. For T1, three classification approaches were considered (Multinomial Naive Bayes, linear SVM with Stochastic Gradient Descent and Passive Aggressive). The team used T1's training data to optimize the classifiers. To meet this aim, they used a bag of words approach and content-based features. For T2, the team used an external depression dataset to build a classifier that was subsequently used to estimate risk of self-harm. For T3, the authors developed a rule-based approach. The 21 depression questions were grouped into 6 categories and, for each category, the estimated responses were produced by combining multiple features such as the presence of certain lexical categories in the user's text, the use of absolutist language, etc.

BiTeM: This team comes from the HES-SO/HEG Geneva and the Swiss Institute of Bioinformatics. They developed several supervised learning solutions for T1 and T2. This included standard SVMs with content-based features, convolutional neural networks and ensemble methods. These participants obtained additional training data from several Reddit communities and extracted post-level annotations using mutual information. For T3, they developed a data-driven, ensemble model approach that leverages word polarities, token extraction via mutual information, keyword expansion and semantic similarities.

LIRMM: This group of researchers comes from LIRMMM at the University of Montpellier (France). This team developed a method based on modeling temporal mood variation. This is a two-stage approach with an initial Deep Mood Module, which employs attention-based deep learning models to build a time series that represents the temporal mood variation found in the user's thread of posts, and a second module that makes the final decision based on machine learning or Bayesian inference. For T2, they obtained training data from the depression and anorexia collections of previous eRisks.

ClaC: This is a team from the Computational Linguistics at Concordia Laboratory (Concordia University, Canada). They developed an ensemble approach that employs several attention-based neural sub-models to extract features and predict class probabilities. These features were later used as input features to a Support Vector Machine (SVM) that made the final estimation. The resulting methods performed very well under most evaluation metrics. Another strong aspect of these algorithms is that they worked with a low number of user posts. This suggests that the proposed solution has potential to make early and correct decisions.

SINAI: These participants come from the University of Jaén (Spain) and designed three supervised learning variants for T1. The first variant was a SVM with tf/idf weights, the second run was a variant of the first run that manipulated the original weights to incorporate the semantic similarity of each word with the word *anorexia* (using similarity embeddings), and the third run considered additional external information, obtained from the Unified Medical Language System (UMLS).

HULAT: This group of researchers are affiliated to Universidad Carlos III (Spain) and implemented five different variants of deep neural networks to detect at-risk users. In their experiments, they considered different types of neural networks such as RNN and CNN, and some submitted variants built on transfer learning elements.

UDE: This is a team composed of researchers from the University of Duisburg-Essen (Germany) and the University of Toulouse (France). They implemented a number of variants ranging from standard Support Vector Machine solutions to more elaborated neural network classifiers. Their runs included a i) linear kernel SVM, built from the training set using content-based features, ii) a two-stage SVM approach, where an initial SVM is built to filter out offtopic posts and, next, another SVM makes the final classification, iii) a Long Short Term Memory neural network model, iv) a global attention deep learning model and v) a inner attention model. Some of these variants required post-level annotations, which were produced by the participants, or external resources, such as webpages from Eating Disorder sites. For T2 (self-harm), this team opted for building predictive models that were trained with depression and anorexia data.

SSN-NLP: This group comes from the SSN College of Engineering, Chennai (India). They employed variations of two major models for sentiment classification: a deep learning RNN-LSTM and a traditional SGDC Classifier. More specifically, they submitted four Deep Learning variants and one traditional learning model that uses Neural Machine Translation (NMT) and SVM with SGD optimizer.

UNSL: This is a joint collaboration between Universidad Nacional de San Luis (Argentina), Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina), and Instituto Nacional de Astrofísica, Óptica y Electrónica (México). This team built on the results they got in previous years and performed experiments with a text classification

approach that supports incremental training, early classification and visual explanations. They employed the same core classifier, SS3, for the three tasks. Essentially, it is a word-based approach that estimates risk based on a number of term statistics (within-class frequency, within-class significance and inter-class term significance). The T2 variants focused on how to produce training data for this task. The T3 experiments focused on how to employ the classification approach to answer the 21 questions in the depression questionnaire.

LTL-INAOE: This team is affiliated to the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico. Their approach is based on identifying personal phrases, which are those where certain personal pronouns occur, and extracting content-based features from those phrases. Terms are selected and weighted based on the distribution of occurrences in personal and non-personal phrases, and a linear SVM is employed for the final prediction. For T2, estimation was done by matching user posts with data retrieved from the self-harm community of Reddit.

INAOE-CIMAT: This is a joint collaboration between Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico and Centro de Investigación en Matemáticas (CIMAT), Mexico. This team employed a supervised learning method with SVMs together with a chi-squared approach for extracting the most relevant features. The base features come from the so-called *BoSE* representation process (Bag of Subemotions). Essentially, seed words are obtained from an emotion lexicon and, next, these words are used to extract fine-grained emotions (with an affinity propagation clustering algorithm). The authors report experiments with the training data that show that the BoSE representation is substantially better than a Bag-of-Words approach.

CAMH: This team, which comes from multiple Canadian institutions, participated in T2 and T3. For T2 (self-harm), they got training data from the University of Maryland (UMD) Reddit Suicidality Dataset and developed a supervised learning approach whose features were extracted using a Language Modeling approach. For T3, they first represented the users with features similar to those used for T2 (together with additional LIWC features) and, next, for each question in the BDI questionnaire, a vectorial representation of the user is matched against a vectorial representation associated to each possible response.

2.4 Task 1: Results

Table 3 shows the participating teams, the number of runs submitted and the approximate lapse of time from the first response to the last response. This lapse of time is indicative of the degree of automation of each team's algorithms. Most of the submitted runs processed the entire thread of messages (around 2000 iterations), but a few variants opted for stopping earlier. Only a few teams (HULAT, BiTeM, BioInfo@UAVR and UNSL) processed the thread of messages in a reasonably fast way (less than a day for processing the entire history of user messages). The rest of the teams took several days to run the whole process. This suggests that they incorporated some form of offline processing.

Table 4 reports the decision-based performance achieved by the participating teams. In terms of $F1$ and latency-weighted $F1$, the best performing run was sent by the ClaC team. The runs submitted by this team suggest that you can get to a level of effectiveness

Table 3. Task 1. Participating teams: number of runs, number of user writings processed by the team, and lapse of time taken for the whole process.

team	#runs	#user writings processed	lapse of time (from 1st to last response)
UppsalaNLP	5	2000	2 days + 7 hs
BioInfo@UAVR	1	2000	14 hs
BiTeM	5	11	4 hs
lirmm	5	2024	8 days + 15 hs
CLaC	5	109	11 days + 16 hs
SINAI	3	317	10 days + 7 hs
HULAT	5	83	18 hs
UDE	5	2000	5 days + 3 hs
SSN-NLP	5	9	6 days + 22 hs
Fazl	3	2001	21 days + 15 hs
UNSL	5	2000	23 hs
LTL-INAOE	2	2001	17 days + 23 hs
INAOE-CIMAT	5	2000	8 days + 2 hs

of about 70% based on a few user writings. As a matter of fact, the best performing run had a median of 7 user writings analyzed for the true positives detected. Other teams submitted quicker decisions ($latency_{TP}$ and $speed$ equal to 1) but the associated effectiveness was poor. This is not surprising because decisions based on a single user post are likely premature.

In terms of precision, the lirmm team sent a run with 77% performance. This variant, which also had reasonable figures for recall and $F1$, looks promising and its true positive decisions were fast (about 20 user posts processed). Some teams submitted runs with very high values of recall but the associated precision and other metrics were very low.

In terms of $ERDE$, the two best performing runs were sent by UNSL. As argued above, this measure sets a strong penalty on late decisions and this teams opted to send true positive decisions after seeing a couple of user writings.

Overall, these results suggest that with a few dozen user writings some systems led to reasonably high effectiveness. The best predictive algorithms could be used to support expert humans in early detecting signs of anorexia.

Table 5 reports the ranking-based performance achieved by the participating teams. Many teams only processed a few dozens of user writings and, thus, we could only compute their rankings of users for the initial points. Other teams (e.g., UppsalaNLP or BioInfo@UAVR) have the same ranking-based effectiveness over multiple points (after 1 writing, after 100 writings, and so forth). This suggests that these teams did not change the risk scores estimated from the initial stages.

Other participants (ClaC, UDE, Fazl, UNSL and LTL-INAOE) behave as expected: the rankings of estimated risk get better as they are built from more user evidence. Notably, some UNSL and UDE variants led to almost perfect $P@10$ and $NDCG@10$ performance after analyzing more than 100 writings. This suggests that, with enough

Table 4. Task 1. Decision-based evaluation

team	run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	*speed*	*latency-weighted* $F1$
UppsalaNLP	0	.32	.44	.37	5.83%	5.77%	**1**	**1**	.37
UppsalaNLP	1	.36	.39	.37	6.13%	6.07%	**1**	**1**	.37
UppsalaNLP	2	.34	.42	.38	5.88%	5.81%	**1**	**1**	.38
UppsalaNLP	3	.39	.30	.34	6.68%	6.63%	**1**	**1**	.34
UppsalaNLP	4	.40	.42	.41	5.73%	5.66%	**1**	**1**	.41
BioInfo@UAVR	0	.32	.44	.37	5.84%	5.77%	**1**	**1**	.37
BiTeM	0	.42	.07	.12	8.58%	8.42%	**1**	**1**	.12
BiTeM	1	.44	.70	.54	5.89%	3.40%	3	.99	.54
BiTeM	2	.73	.11	.19	8.42%	8.01%	3	.99	.19
BiTeM	3	1	.01	.03	8.84%	8.83%	1	1	.03
BiTeM	4	0	0	0	-	-	-	-	-
lirmm	0	.74	.63	.68	9.13%	5.14%	21	.92	.63
lirmm	1	**.77**	.60	.68	9.10%	5.51%	21	.92	.62
lirmm	2	.66	.70	.68	9.24%	5.81%	31	.88	.60
lirmm	3	.74	.42	.54	9.08%	6.62%	31	.88	.48
lirmm	4	.57	.75	.65	9.41%	7.32%	2023	$3e^{-7}$	$2e^{-7}$
CLaC	0	.45	.74	.56	6.72%	3.93%	7	.98	.54
CLaC	1	.61	.82	.70	5.73%	3.13%	4	.99	.69
CLaC	2	.60	.81	.69	6.02%	3.13%	6	.98	.68
CLaC	3	.63	.76	.69	6.27%	3.55%	7	.98	.68
CLaC	4	.64	.79	**.71**	6.25%	3.43%	7	.98	**.69**
SINAI	0	.12	.97	.21	10.58%	6.59%	5	.98	.21
SINAI	1	.11	.99	.20	10.80%	6.76%	5	.98	.20
SINAI	2	.18	.95	.30	9.04%	4.89%	8	.97	.30
HULAT	0	.11	.30	.17	10.84%	8.14%	16.5	.94	.16
HULAT	1	.11	.30	.17	10.84%	8.14%	16.5	.94	.16
HULAT	2	.11	.30	.17	10.84%	8.14%	16.5	.94	.16
HULAT	3	.11	.30	.17	10.84%	8.14%	16.5	.94	.16
HULAT	4	.11	.30	.17	10.84%	8.14%	16.5	.94	.16
UDE	0	.51	.74	.61	8.48%	3.87%	11	.96	.58
UDE	1	.44	.73	.55	7.48%	3.94%	9	.97	.53
UDE	2	.13	.68	.22	12.52%	8.21%	35	.87	.19
UDE	3	0	0	0	-	-	-	-	-
UDE	4	0	0	0	-	-	-	-	-
SSN-NLP	0	.32	.16	.22	8.24%	7.76%	2	1	.22
SSN-NLP	1	.30	.22	.25	7.90%	7.41%	**1**	**1**	.25
SSN-NLP	2	.47	.22	.30	7.80%	7.19%	2	1	.30
SSN-NLP	3	.48	.26	.34	7.61%	6.86%	2	1	.33
SSN-NLP	4	.32	.15	.21	8.08%	7.86%	**1**	**1**	.21
Fazl	0	.09	**1**	.16	17.11%	13.91%	97	.64	.11
Fazl	1	.09	**1**	.16	17.11%	13.79%	88	.67	.11
Fazl	2	.09	**1**	.16	17.11%	11.22%	34	.87	.14
UNSL	0	.42	.78	.55	**5.54%**	3.92%	2	1	.55
UNSL	1	.43	.75	.55	5.68%	4.10%	2	1	.55
UNSL	2	.36	.86	.51	5.56%	3.34%	2	1	.50
UNSL	3	.35	.85	.50	5.59%	3.49%	2	1	.49
UNSL	4	.31	.92	.47	6.14%	**2.97%**	3	.99	.46
LTL-INAOE	0	.45	.75	.57	7.78%	4.23%	11	.96	.54
LTL-INAOE	1	.47	.75	.58	7.74%	4.20%	11	.96	.55
INAOE-CIMAT	0	.56	.78	.66	9.30%	3.98%	15	.95	.62
INAOE-CIMAT	1	0	0	0	-	-	-	-	-
INAOE-CIMAT	2	.58	.77	.66	9.28%	9.16%	65	.76	.50
INAOE-CIMAT	3	.67	.68	.68	9.17%	4.75%	20	.93	.63
INAOE-CIMAT	4	.69	.63	.66	9.13%	5.08%	20	.93	.61

pieces of evidence, the methods implemented by these teams are highly effective at prioritizing at-risk users.

Table 5. Task 1. Ranking-based evaluation

team	run	1 writing			100 writings			500 writings			1000 writings		
		P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100
UppsalaNLP	0	.6	.59	.47	.6	.59	.47	.6	.59	.47	.6	.59	.47
UppsalaNLP	1	.4	.31	.40	.4	.31	.40	.4	.31	.40	.4	.31	.40
UppsalaNLP	2	.5	.38	.42	.5	.38	.42	.5	.38	.42	.5	.38	.42
UppsalaNLP	3	.7	.65	.45	.7	.65	.45	.7	.65	.45	.7	.65	.45
UppsalaNLP	4	.8	.75	.52	.8	.75	.52	.8	.75	.52	.8	.75	.52
BioInfo@UAVR	0	.6	.59	.47	.6	.59	.47	.6	.59	.47	.6	.59	.47
BiTeM	0	.6	.44	.52	-	-	-	-	-	-	-	-	-
BiTeM	1	.8	.75	.47	-	-	-	-	-	-	-	-	-
BiTeM	2	.8	.71	.46	-	-	-	-	-	-	-	-	-
BiTeM	3	.8	.71	.48	-	-	-	-	-	-	-	-	-
BiTeM	4	.8	.71	.48	-	-	-	-	-	-	-	-	-
lirmm	0	-	-	-	-	-	-	-	-	-	-	-	-
lirmm	1	-	-	-	-	-	-	-	-	-	-	-	-
lirmm	2	-	-	-	-	-	-	-	-	-	-	-	-
lirmm	3	-	-	-	-	-	-	-	-	-	-	-	-
lirmm	4	-	-	-	-	-	-	-	-	-	-	-	-
CLaC	0	.1	.10	.05	.8	.86	.28	-	-	-	-	-	-
CLaC	1	.1	.10	.04	.3	.45	.16	-	-	-	-	-	-
CLaC	2	-	-	-	-	-	-	-	-	-	-	-	-
CLaC	3	-	-	-	-	-	-	-	-	-	-	-	-
CLaC	4	-	-	-	-	-	-	-	-	-	-	-	-
SINAI	0	.2	.12	.11	-	-	-	-	-	-	-	-	-
SINAI	1	.2	.12	.11	-	-	-	-	-	-	-	-	-
SINAI	2	.2	.12	.11	-	-	-	-	-	-	-	-	-
HULAT	0	.3	.33	.18	-	-	-	-	-	-	-	-	-
HULAT	1	.3	.33	.18	-	-	-	-	-	-	-	-	-
HULAT	2	.3	.33	.18	-	-	-	-	-	-	-	-	-
HULAT	3	.3	.33	.18	-	-	-	-	-	-	-	-	-
HULAT	4	.3	.33	.18	-	-	-	-	-	-	-	-	-
UDE	0	.2	.12	.11	.9	.92	.81	.9	.93	.85	.9	.94	.86
UDE	1	.6	.75	.54	.9	.94	.81	1	1	.87	1	1	.88
UDE	2	.7	.76	.49	.9	.94	.60	.9	.94	.64	.8	.88	.64
UDE	3	-	-	-	-	-	-	-	-	-	-	-	-
UDE	4	.0	.0	.11	.0	.0	.08	.0	.0	.06	.0	.0	.07
SSN-NLP	0	.6	.64	.29	-	-	-	-	-	-	-	-	-
SSN-NLP	1	.3	.28	.15	-	-	-	-	-	-	-	-	-
SSN-NLP	2	.5	.48	.29	-	-	-	-	-	-	-	-	-
SSN-NLP	3	.6	.64	.30	-	-	-	-	-	-	-	-	-
SSN-NLP	4	.3	.33	.15	-	-	-	-	-	-	-	-	-
Fazl	0	.2	.12	.11	.1	.10	.26	.0	.0	.35	.1	.06	.39
Fazl	1	.3	.29	.26	.6	.60	.59	.7	.78	.67	.7	.78	.68
Fazl	2	.2	.12	.11	.8	.82	.46	.9	.94	.62	1	1	.66
UNSL	0	.8	.82	.54	1	1	.77	1	1	.79	1	1	.79
UNSL	1	.8	.82	.54	1	1	.77	1	1	.79	1	1	.79
UNSL	2	.8	.82	.55	1	1	.83	1	1	.83	1	1	.84
UNSL	3	.8	.82	.53	1	1	.83	1	1	.84	1	1	.84
UNSL	4	.8	.82	.52	.9	.94	.85	1	1	.85	.9	.94	.84
LTL-INAOE	0	.8	.75	.34	1	1	.76	.9	.92	.73	.7	.78	.65
LTL-INAOE	1	.8	.75	.34	1	1	.76	.9	.92	.73	.7	.78	.66
INAOE-CIMAT	0	-	-	-	-	-	-	-	-	-	-	-	-
INAOE-CIMAT	1	-	-	-	-	-	-	-	-	-	-	-	-
INAOE-CIMAT	2	-	-	-	-	-	-	-	-	-	-	-	-
INAOE-CIMAT	3	-	-	-	-	-	-	-	-	-	-	-	-
INAOE-CIMAT	4	-	-	-	-	-	-	-	-	-	-	-	-

3 Task 2: Early Detection of Signs of Self-harm

This task is new in 2019. T2 has a similar organization as T1, but T2 provided no training data. The challenge consists of sequentially processing pieces of evidence and detect early traces of self-harm as soon as possible. There are two categories of users, self-harm and non-self-harm, and, for each user, the collection contains a sequence of writings (in chronological order). T2 had only a test stage (no training stage) and, therefore, we encouraged participants to design their own unsupervised (e.g. search-based) strategies to detect possible cases of self-harm. Similar to T1, the test stage

Table 6. Task2 (self-harm). Main statistics of the collection

	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

consisted of a period of time where the participants had to connect to our server and iteratively get user writings and send responses.

Table 6 reports the main statistics of the T2 dataset. The self-harm group is composed of users who were active on the self-harm Reddit community and explicitly said that they had committed self-harm (e.g., cuts or injuries). We wanted to further instigate the creation of algorithms that detect these cases as early as possible. To meet this aim, for each individual, the algorithms were given only the history of the postings *before* the individual entered into the self-harm community (i.e. all posts before the first entry in the self-harm communities). We thought that an individual who is active on self-harm forums perhaps has already done some sort of self-harm to his body and 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 were only given the texts posted by the affected individuals before they first engaged in the self-harm community. Similar to T1, the systems had an item-by-item access to the user's history of posts.

We expected that effectiveness was lower compared to T1. First, because T2 provided no training data. Second, because user history consisted solely on the postings before entering the self-harm community (and, thus, signals related to self-harm might not be that explicit).

The evaluation approach employed for T2 was exactly the same used for T1 (see section 2).

3.1 Task 2: Results

Table 7 shows the participating teams, the number of runs submitted, and the approximate lapse of time from the first response to the last response. Most of the submitted runs processed the entire thread of postings (around 2000 iterations), except for one participant (BiTeM) that opted for stopping earlier. Compared with T1, the teams were quicker at processing the entire thread of user writings but there were still some teams that took more than a day for running the whole estimation process. Again, this suggests that some participants incorporated some form of offline processing.

Table 8 reports the decision-based metrics. Not surprisingly, effectiveness scores are lower than those achieved for T1. $F1$ and its latency-weighted version are barely higher than 50% for the best performing runs. The best performing run, from UNSL, was extremely fast at making decisions (the median number of postings analyzed before making a true positive decision was about 2) but the effectiveness of its decisions was rather modest. A few runs deeply analyzed the entire user history (e.g., some runs from

Table 7. Task 2 (Self-harm). Participating teams: number of runs, number of user writings processed by the team, and lapse of time taken for the whole process.

team	#runs	#user writings processed	lapse of time (from 1st to last response)
BiTeM	5	8	3 min
BioInfo@UAVR	1	1992	4 hs
Fazl	3	1993	18 days + 21 hs
UNSL	5	1992	13 hs
UDE	5	1992	1 day + 2 hs
LTL-INAOE	4	1993	17 hs
lirmm	5	2004	2 days + 22 hs
CAMH	5	1992	1 day + 19 hs

lirmm) but this did not lead to better decisions. Overall, these results suggest that it is unclear how early traces of self-harm could be detected from the user interactions in Social Media prior to their first entry in the self-harm community. In the future, it would be interesting to study how much benefit these algorithms can receive from training data.

The corresponding ranking-based scores are reported in Table 9. The *initial* ranking-based performance (rankings based on a single writing) is low for most of the participants. However, some runs (particularly some UNSL runs) managed to produce effective rankings after analyzing 100 or more user posts. This suggests that the tendency to make early alerts (as indicated by the latency and speed statistics shown in Table 8) was detrimental to the identification of at-risk individuals.

4 Task 3: Measuring the Severity of the Signs of Depression

This is a new task in 2019. The task consists of estimating the level of depression from a thread of user submissions. For each user, the participants were given its full history of postings (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 questionnaires are derived from the Beck's Depression Inventory (BDI)[1], which assesses the presence of feelings like sadness, pessimism, loss of energy, etc. for the detection of depression. The questionnaire contains 21 questions (see figs 2, 3).

The task aims 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 participating systems.

The participants were given a dataset with 20 users and they were asked to produce a file with the following structure:

```
username1 answer1 answer2 .... answer21
```

Instructions:

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel. If several statements in the group seem to apply equally well, choose the highest number for that group.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don't feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

6. Punishment Feelings
0. I don't feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.

7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.

8. Self-Criticalness
0. I don't criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes
0. I don't have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying
0. I don't cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can't.

11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it's hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It's hard to get interested in anything.

13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness
0. I do not feel I am worthless.
1. I don't consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don't have enough energy to do very much.
3. I don't have enough energy to do anything.

Fig. 2. Beck's Depression Inventory (part 1)

```
16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
1a. I sleep somewhat more than usual.
1b. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a lot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.

18. Changes in Appetite
0. I have not experienced any change in my appetite.
1a. My appetite is somewhat less than usual.
1b. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.

19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can't concentrate as well as usual.
2. It's hard to keep my mind on anything for very long.
3. I find I can't concentrate on anything.

20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely
```

Fig. 3. Beck's Depression Inventory (part 2)

```
username2 ....
....
```

Each line has a user identifier and 21 values. These values correspond to the responses to the questions of the depression questionnaire (the possible values are 0, 1a, 1b, 2a, 2b, 3a, 3b -for questions 16 and 18- and 0, 1, 2, 3 -for the rest of the questions-).

4.1 Task 3: Evaluation Metrics

We considered a number of metrics in order to assess the quality of a questionnaire filled by a system when compared to the real questionnaire filled by actual Social Media user:

- **Average Hit Rate** (AHR): Hit Rate (HR) averaged over all users. HR is a stringent measure that computes the ratio of cases where the automatic questionnaire has exactly the same answer as the real questionnaire. For example, an automatic questionnaire with 5 matches gets HR equal to 5/21 (because there are 21 questions in the form).
- **Average Closeness Rate** (ACR): Closeness Rate (CR) averaged over all users. CR takes into account that the answers of the depression questionnaire represent an ordinal scale. For example, consider the #17 question:

```
17. Irritability
```

```
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.
```

Imagine that the real user answered "0". A system S1 whose answer is "3" should be penalised more than a system S2 whose answer is "1".
For each question, CR computes the absolute difference (ad) between the real and the automated answer (e.g. ad=3 and ad=1 for S1 and S2, respectively) and, next, this absolute difference is transformed into an effectiveness score as follows: $CR = (mad - ad)/mad$, where mad is the maximum absolute difference, which is equal to the number of possible answers minus one.
NOTE: in the two questions (#16 and #18) that have seven possible answers $\{0, 1a, 1b, 2a, 2b, 3a, 3b\}$ the pairs $(1a, 1b)$, $(2a, 2b)$, $(3a, 3b)$ are considered equivalent because they reflect the same depression level. As a consequence, the difference between $3b$ and 0 is equal to 3 (and the difference between $1a$ and $1b$ is equal to 0).

- **Average DODL (ADODL)**: Difference between overall depression levels (DODL) averaged over all users. The previous measures assess the systems' ability to answer each question in the form. DODL, instead, does not look at question-level hits or differences but computes the overall depression level (sum of all the answers) for the real and automated questionnaire and, next, the absolute difference ($ad\_overall$) between the real and the automated score is computed.
  Depression levels are integers between 0 and 63 and, thus, DODL is normalised into [0,1] as follows: $DODL = (63 - ad\_overall)/63$.
- **Depression Category Hit Rate (DCHR)**. In the psychological domain, it is customary to associate depression levels with the following categories:

```
minimal depression (depression levels 0-9)
mild depression (depression levels 10-18)
moderate depression (depression levels 19-29)
severe depression (depression levels 30-63)
```

  The last effectiveness measure consists of computing the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire.

4.2 Task 3: Results

Table 10 (upper block) presents the results achieved by the participants in this task. To put things in perspective, the table also reports (lower block) the performance achieved by three baseline variants: *all 0s* and *all 1s*, which consist of sending the same response (0 or 1) for all the questions, and *random*, which is the average performance (averaged over 1000 repetitions) achieved by an algorithm that randomly chooses among the possible answers.

In terms of AHR, the best performing run (ANSLC) shows that it is possible to get more than 40% of the answers right. The distance-based variant (ACR) shows some

figures greater than 70% (e.g. for UNSLE). These performance metrics are higher than those achieved by a random algorithm. This suggests that the analysis of the user posts is useful at extracting some signals or symptoms related to depression. However, the näive *all 1s* algorithm yields to the highest ACR score. This metric penalizes high distances between the correct answer and the answer given by the system and, thus, it somehow favours *conservative* answers. By choosing always 1, the *all 1s* algorithm sets an upper limit of the distance equal to 2 (it gets 2 when the correct answer is 3) and, in our comparison, it becomes the top performing choice. However, some participating runs are clearly better than the three baseline algorithms in terms of AHR. This suggests that the teams do a reasonable job at getting answers right but, when they fail, the distance-based ACR scores penalizes them severely.

ADODL and, particularly, DCHR show that the participants, although effective at answering some depression-related questions, do not fare well at estimating the overall level of depression of the individuals. For example, the best performing run gets the depression category right for only 45% of the individuals. This is better than the baseline variants but, still, there is much room for improvement.

Overall, these experiments indicate that it is possible to automatically extract some depression-related evidence from social media activity but we are still far from a really effective depression screening tool. In the near future, it will be interesting to further analyze the participants' estimations in order to investigate which particular BDI questions are easier or harder to automatically answer based on Social Media activity.

5 Conclusions

This paper provided an overview of eRisk 2019. This was the third edition of this lab and the lab's activities concentrated on two different types of tasks: early detection of signs of anorexia (T1) or self-harm (T2), where the participants had a sequential access to the user's social media posts and they had to send alerts about at-risk individuals, and measuring the severity of the signs of depression (T3), where the participants were given the full user history and their systems had to automatically estimate the user's responses to a standard depression questionnaire.

Overall, the proposed tasks received 105 variants or runs from 14 participants. Although the effectiveness of the proposed solutions is still modest, the experiments suggest that evidence extracted from Social Media is valuable and automatic or semi-automatic screening tools could be designed to detect at-risk individuals. This promising result encourages us to further explore the creation of benchmarks for text-based screening of signs of risk.

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, Universi-

dade 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 (06 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: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016. Evora, Portugal (2016)
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., 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)
5. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2017. Dublin, Ireland (2017)
6. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRISK 2018: Early Risk Prediction on the Internet (extended lab overview). In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2018. Avignon, France (2018)
7. 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)
8. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: WSDM. pp. 495–503. ACM (2018)
9. 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)

Table 8. Task 2 (Self-harm). Decision-based evaluation

team	run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
BiTeM	0	.52	.41	.46	9.73%	7.63%	3	.99	.46
BiTeM	1	1	.05	.09	11.84%	11.47%	6.5	.98	.09
BiTeM	2	0	0	0	-	-	-	-	-
BiTeM	3	0	0	0	-	-	-	-	-
BiTeM	4	0	0	0	-	-	-	-	-
BioInfo@UAVR	0	.55	.39	.46	10.79%	8.11%	6	.98	.45
Fazl	0	.12	**1**	.22	22.66%	16.71%	51	.81	.17
Fazl	1	.12	**1**	.22	22.66%	16.42%	47	.82	.18
Fazl	2	.12	**1**	.22	22.66%	13.24%	35	.87	.19
UNSL	0	**.71**	.41	**.52**	9.01%	7.31%	2	1	**.52**
UNSL	1	.70	.39	.50	9.03%	7.60%	2.5	.99	.50
UNSL	2	.20	.90	.32	9.20%	6.86%	2	1	.32
UNSL	3	.31	.85	.45	8.79%	5.44%	3	.99	.45
UNSL	4	.31	.88	.46	**8.21%**	**4.93%**	3	.99	.45
UDE	0	.50	.07	.13	12.02%	11.28%	13	.95	.12
UDE	1	.45	.22	.30	11.27%	9.80%	7	.98	.29
UDE	2	.18	.68	.29	13.75%	9.56%	13.5	.95	.28
UDE	3	.06	.34	.10	20.04%	20.02%	73.5	.72	.07
UDE	4	0	0	0	-	-	-	-	-
LTL-INAOE	0	.12	**1**	.22	12.53%	10.60%	**1**	**1**	.22
LTL-INAOE	1	.12	**1**	.22	13.15%	10.90%	2	1	.21
LTL-INAOE	2	.12	**1**	.22	16.64%	10.90%	4	.99	.21
LTL-INAOE	3	.12	**1**	.22	17.15%	10.90%	5	.98	.21
lirmm	0	.57	.29	.39	12.22%	10.02%	28.5	.89	.35
lirmm	1	.53	.22	.31	12.18%	10.58%	21	.92	.29
lirmm	2	.48	.49	.48	12.84%	11.67%	2004	$3e^{-7}$	$1e^{-7}$
lirmm	3	.47	.44	.46	12.77%	11.89%	2004	$3e^{-7}$	$1e^{-7}$
lirmm	4	.52	.41	.46	12.63%	11.74%	2004	$3e^{-7}$	$1e^{-7}$
CAMH	0	.12	.95	.22	16.67%	10.60%	7	.98	.22
CAMH	1	.12	.93	.22	16.81%	10.98%	7	.98	.21
CAMH	2	.12	.90	.22	17.41%	11.49%	8	.97	.21
CAMH	3	.12	.98	.22	15.70%	10.70%1	3.5	.99	.22
CAMH	4	.12	**1**	.22	14.84%	10.32%	4	.99	.22

Table 9. Task 2 (Self-harm). Ranking-based evaluation

team	run	1 writing			100 writings			500 writings			1000 writings		
		P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100	P@10	$NDCG$ @10	$NDCG$ @100
BiTeM	0	.3	.35	.53	-	-	-	-	-	-	-	-	-
BiTeM	1	.4	.47	.39	-	-	-	-	-	-	-	-	-
BiTeM	2	.5	.48	.44	-	-	-	-	-	-	-	-	-
BiTeM	3	.2	.38	.41	-	-	-	-	-	-	-	-	-
BiTeM	4	.4	.56	.50	-	-	-	-	-	-	-	-	-
BioInfo@UAVR	0	-	-	-	-	-	-	-	-	-	-	-	-
Fazl	0	.1	.12	.30	.1	.06	.42	.1	.06	.41	.6	.40	.59
Fazl	1	.2	.27	.36	.9	.94	.83	.9	.94	.84	.9	.94	.84
Fazl	2	0	0	.13	.6	.73	.73	.7	.68	.71	.7	.68	.74
UNSL	0	.7	.79	.48	.9	.94	.61	.9	.94	.66	.9	.94	.66
UNSL	1	.6	.74	.48	.9	.94	.60	.9	.94	.65	.9	.94	.65
UNSL	2	.9	.88	.62	.8	.75	.75	.5	.59	.74	.6	.64	.74
UNSL	3	1	1	.67	.9	.94	.84	.7	.63	.75	.7	.63	.75
UNSL	4	1	1	.64	.9	.93	.86	.7	.67	.79	.8	.74	.78
UDE	0	0	0	.09	.7	.77	.69	.7	.67	.69	.7	.67	.70
UDE	1	.7	.56	.52	.7	.66	.69	.8	.75	.74	.8	.75	.74
UDE	2	.5	.63	.53	.5	.56	.64	.6	.66	.68	.6	.65	.67
UDE	3	.2	.19	.21	0	0	.11	.2	.16	.14	.1	.07	.15
UDE	4	.2	.25	.30	.1	.07	.20	.1	.07	.15	.1	.08	.17
LTL-INAOE	0	.5	.50	.45	.4	.41	.55	.2	.19	.25	.1	.19	.37
LTL-INAOE	1	.6	.73	.56	.1	.19	.17	.1	.06	.07	0	0	.04
LTL-INAOE	2	.4	.42	.32	.1	.07	.43	.2	.19	.25	.1	.19	.37
LTL-INAOE	3	.7	.72	.44	.1	.19	.27	.1	.06	.19	0	0	.29
lirmm	0	.1	.19	.15	0	0	.01	-	-	-	-	-	-
lirmm	1	.1	.19	.15	0	0	.01	-	-	-	-	-	-
lirmm	2	.1	.19	.15	0	0	.01	-	-	-	-	-	-
lirmm	3	.1	.19	.15	0	0	.01	-	-	-	-	-	-
lirmm	4	.1	.19	.15	0	0	.01	-	-	-	-	-	-
CAMH	0	.3	.37	.47	.6	.71	.49	.7	.72	.50	.6	.66	.48
CAMH	1	.4	.41	.43	.6	.65	.42	.7	.72	.49	.6	.66	.47
CAMH	2	.3	.41	.51	.5	.62	.39	.7	.72	.45	.6	.66	.44
CAMH	3	.3	.25	.42	.5	.55	.32	.7	.72	.37	.6	.66	.37
CAMH	4	.5	.48	.50	.6	.59	.34	.6	.66	.43	.6	.66	.39

Table 10. Task 3. Performance Results

Run	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	34.05%	66.43%	77.70%	25.00%
BiTeM	32.14%	62.62%	72.62%	25.00%
CAMH_GPT_nearest_unsupervised	23.81%	57.06%	**81.03%**	**45.00%**
CAMH_GPT_supervised.181_features.58hr	35.47%	68.33%	75.63%	20.00%
CAMH_GPT_supervised.769_features.55hr	36.43%	67.22%	72.30%	20.00%
CAMH_GPT_supervised.949_features.75hr	36.91%	69.13%	75.63%	15.00%
CAMH_LIWC_supervised_SVM	35.95%	66.59%	75.48%	25.00%
Fazl	22.38%	56.27%	72.78%	5.00%
Illinois	22.62%	56.19%	66.35%	40.00%
ISIKol_multiSimilarity-5000-Dtac-Qtac	29.76%	57.94%	74.13%	25.00%
ISIKol-bm25-1.2-0.75-5000-Dtac-Qtac	29.76%	57.06%	72.78%	25.00%
ISIKol-lm-d-1.0-5000-Dtac-Qtac	30.00%	57.94%	73.02%	15.00%
Kimberly	38.33%	64.44%	66.19%	20.00%
UNSLA	37.38%	67.94%	72.86%	30.00%
UNSLB	36.93%	70.16%	76.83%	30.00%
UNSLC	**41.43%**	69.13%	78.02%	40.00%
UNSLD	38.10%	67.22%	78.02%	30.00%
UNSLE	40.71%	**71.27%**	80.48%	35.00%
all 0s	34.52%	61.43%	61.43%	20.00%
all 1s	27.14%	71.75%	79.37%	20.00%
random (avg 1000 repetitions)	23.98%	58.55%	77.78%	33.55%