eRisk 2019: T1 results

No Author Given

No Institute Given

1 Task 1: Early Detection of Signs of Anorexia

This is a 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 created. 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 for this task had the same format as the collection described in [1]. The source of data is also the same used for eRisk 2017 and 2018. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, anorexia and non-anorexia, and, for each user, the collection contains a sequence of writings (in chronological order).

In 2019, we moved from a chunk-based release of data (used in 2017 and 2018) to a item-by-item release of data. We set up a server that iteratively gave user writings to the participating teams. More information about the server can be found at the lab website[1].

2 Task 1: evaluation metrics

2.1 Decision-based evaluation

The evaluation of the submitted runs considered $ERDE$, the early risk detection measure proposed in [1], but it also considered alternative decision-based metrics.

$ERDE$ has a number of drawbacks, namely:

- the penalty associated to true positives goes quickly to 1. This is because of 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 (users with few writings per chunk vs users with many writings per chunk).
- $ERDE$ is not interpretable.

[1] http://early.irlab.org/server.html

Some research teams have analysed these issues and proposed alternative ways for evaluation. More specifically, Trotzek and colleagues [3] 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 (now, all users weight the same). However, there is an important limitation of $ERDE_o^{\%}$. In practice, in real life applications, the overall number of user writings is not known in advance (users post contents in Social Media and you have to make predictions with the evidence seen, you do not know when (and if) a user's thread of message is exhausted). Thus, the performance metric should not depend on knowledge about the total number of user writings.

Another proposal of an alternative evaluation metric for early risk prediction was done by Sadeque and colleagues [2]. 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 estimation 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 that detect these cases too late). Therefore, a first and intuitive measure of delay can be defined as follows[2]:

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

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[3].

[2] Observe that Sadeque et al (see [2], 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.

[3] Again, we adopt Sadeque et al.'s proposal but we estimate latency only over the true positives.

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:

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

p is a parameter that determines how quickly the penalty should increase. In [2], p was set such that the penalty equals 0.5 at the median number of posts of a user[4]. 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.

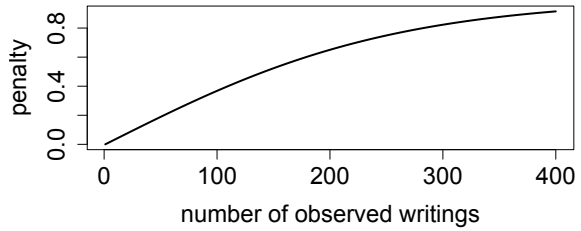

Figure 1. Latency penalty increases with the number of observed writings (k_u)

The system's overall speed factor is computed as:

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

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:

$$F_{latency} = F \cdot speed \tag{7}$$

In 2019, user's data was processed by the participants in a writing by writing 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.

[4] In the eRisk 2017 collection this led to setting p to 0.0078.

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

3 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 (when participants get one writing per user) 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 (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).

4 Task 1: results

Table 1 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 and BioInfo@UAVR) 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.

References

1. David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In *Proceedings Conference and Labs of the Evaluation Forum CLEF 2016*, Evora, Portugal, 2016.
2. Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the latency of depression detection in social media. In *WSDM*, pages 495–503. ACM, 2018.
3. Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *CoRR*, abs/1804.07000, 2018.

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

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

team	run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UppsalaNLP	0	.32	.44	.37	.06	.06	**1**	**1**	.37
UppsalaNLP	1	.36	.39	.37	.06	.06	**1**	**1**	.37
UppsalaNLP	2	.34	.42	.38	.06	.06	**1**	**1**	.38
UppsalaNLP	3	.39	.30	.34	.07	.07	**1**	**1**	.34
UppsalaNLP	4	.40	.42	.41	.06	.06	**1**	**1**	.41
BioInfo@UAVR	0	.32	.44	.37	.06	.06	**1**	**1**	.37
BiTeM	0	.42	.07	.12	.09	.08	**1**	**1**	.12
BiTeM	1	.44	.70	.54	.06	.03	3	.99	.54
BiTeM	2	.73	.11	.19	.08	.08	3	.99	.19
BiTeM	3	1	.01	.03	.09	.09	1	1	.03
BiTeM	4	0	0	0	-	-	-	-	-
lirmm	0	.74	.63	.68	.09	.05	21	.92	.63
lirmm	1	**.77**	.60	.68	.09	.06	21	.92	.62
lirmm	2	.66	.70	.68	.09	.06	31	.88	.60
lirmm	3	.74	.42	.54	.09	.07	31	.88	.48
lirmm	4	.57	.75	.65	.09	.07	2023	$3e^{-7}$	$2e^{-7}$
CLaC	0	.45	.74	.56	.07	.04	7	.98	.54
CLaC	1	.61	.82	.70	.06	.03	4	.99	.69
CLaC	2	.60	.81	.69	.06	.03	6	.98	.68
CLaC	3	.63	.76	.69	.06	.04	7	.98	.68
CLaC	4	.64	.79	**.71**	.06	.03	7	.98	**.69**
SINAI	0	.12	.97	.21	.11	.07	5	.98	.21
SINAI	1	.11	.99	.20	.11	.07	5	.98	.20
SINAI	2	.18	.95	.30	.09	.05	8	.97	.30
HULAT	0	.11	.30	.17	.11	.08	16.5	.94	.16
HULAT	1	.11	.30	.17	.11	.08	16.5	.94	.16
HULAT	2	.11	.30	.17	.11	.08	16.5	.94	.16
HULAT	3	.11	.30	.17	.11	.08	16.5	.94	.16
HULAT	4	.11	.30	.17	.11	.08	16.5	.94	.16
UDE	0	.51	.74	.61	.08	.04	11	.96	.58
UDE	1	.44	.73	.55	.07	.04	9	.97	.53
UDE	2	.13	.68	.22	.13	.08	35	.87	.19
UDE	3	0	0	0	-	-	-	-	-
UDE	4	0	0	0	-	-	-	-	-
SSN-NLP	0	.32	.16	.22	.08	.08	2	1	.22
SSN-NLP	1	.30	.22	.25	.08	.07	**1**	**1**	.25
SSN-NLP	2	.47	.22	.30	.08	.07	2	1	.30
SSN-NLP	3	.48	.26	.34	.08	.07	2	1	.33
SSN-NLP	4	.32	.15	.21	.08	.08	**1**	**1**	.21
Fazl	0	.09	**1**	.16	.17	.14	97	.64	.11
Fazl	1	.09	**1**	.16	.17	.14	88	.67	.11
Fazl	2	.09	**1**	.16	.17	.11	34	.87	.14
UNSL	0	.42	.78	.55	**.06**	.04	2	1	.55
UNSL	1	.43	.75	.55	.06	.04	2	1	.55
UNSL	2	.36	.86	.51	.06	.03	2	1	.50
UNSL	3	.35	.85	.50	.06	.03	2	1	.49
UNSL	4	.31	.92	.47	.06	**.03**	3	.99	.46
LTL-INAOE	0	.45	.75	.57	.08	.04	11	.96	.54
LTL-INAOE	1	.47	.75	.58	.08	.04	11	.96	.55
INAOE-CIMAT	0	.56	.78	.66	.09	.04	15	.95	.62
INAOE-CIMAT	1	0	0	0	-	-	-	-	-
INAOE-CIMAT	2	.58	.77	.66	.09	.09	65	.76	.50
INAOE-CIMAT	3	.67	.68	.68	.09	.05	20	.93	.63
INAOE-CIMAT	4	.69	.63	.66	.09	.05	20	.93	.61

Table 2. Decision-based evaluation

		1 writing			100 writings			500 writings			1000 writings		
team	run	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	-	-	-	-	-	-	-	-	-	-	-	-

Table 3. Ranking-based evaluation