CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations

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

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

Abstract. This paper provides an overview of eRisk 2017. This was the first 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 first edition of eRisk had two possible ways to participate: a pilot task on early risk detection of depression, and a workshop open to the submission of papers related to the topics of the lab.

1 Introduction

The main goal of eRisk was to instigate discussion on the creation of reusable benchmarks for evaluating early risk detection algorithms, by exploring issues of evaluation methodology, effectiveness metrics and other processes related to the creation of test collections for early risk detection. Early detection technologies can be employed in different areas, particularly those related to health and safety. For instance, early alerts could be sent when a predator starts interacting with a child for sexual purposes, or when a potential offender starts publishing antisocial threats on a blog, forum or social network. eRisk wants 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.

Early risk prediction is a challenging and increasingly important research area. However, this area lacks systematic experimental foundations. It is therefore difficult to compare and reproduce experiments done with predictive algorithms running under different conditions.

Citizens worldwide are exposed to a wide range of risks and threats and many of these hazards are reflected on the Internet. Some of these threats stem from criminals such as stalkers, mass killers or other offenders with sexual, racial, religious or culturally related motivations. Other worrying threats might even come from the individuals themselves. For instance, depression may lead to an eating disorder such as anorexia or even to suicide.

In some of these cases early detection and appropriate action or intervention could reduce or minimise these problems. However, the current technology employed to deal with these issues is essentially reactive. For instance, some specific types of risks can be detected by tracking Internet users, but alerts are triggered when the victim makes his disorders explicit, or when the criminal or offending activities are actually happening. We argue that we need to go beyond this late detection technology and foster research on innovative early detection solutions able to identify the states of those at risk of becoming perpetrators of socially destructive behaviour, and the states of those at risk of becoming victims. Thus, we also want to stimulate the development of algorithms that computationally encode the process of becoming an offender or a victim.

It has been shown that the words people use can reveal important aspects of their social and psychological worlds [16]. There is substantial evidence linking natural language to personality, social and situational fluctuations. This is of particular interest to understand the onset of a risky situation and how it reflects the linguistic style of the individuals involved. However, a major hurdle that has to be overcome is the lack of evaluation methodologies and test collections for early risk prediction. In this lab we intended to take the first steps towards filling this gap. We understand that there are two main classes of early risk prediction:

- **Multiple actors**. We include in the first category cases where there is an external actor or intervening factor that explicitly causes or stimulates the problem. For instance, sexual offenders use deliberate tactics to contact vulnerable children and engage them in sexual exploitation. In such cases, early warning systems need to analyse the interactions between the offender and the victim and, in particular, the language of both. The process of predation is known to happen in five phases [11], namely: gaining access, deceptive trust development, grooming, isolation, and approach. Therefore, systems can potentially track conversations and alert about the onset of a risky situation. Initiatives such as the organisation of a sexual predation identification challenge in CLEF [6] have fostered research on mining conversations and identifying predatory behaviour. However, the focus was on identifying sexual predators and predatory text. There was no notion of early warning. We believe that predictive algorithms such as those developed under this challenge [13] could be further evaluated from an early risk prediction perspective. Another example of risk provoked by external actions is terrorist recruitment. There is currently massive online activity aiming at recruiting young people –particularly, teenagers– for joining criminal networks. Excellent work in this area has been done by the AI Lab of the University of Arizona. Among many other things, this team has created a research infrastructure called “the Dark Web” [19], that is available to social science researchers, computer and information scientists, and policy and security analysts. It permits to study a wide range of social and organizational phenomena of criminal networks. The Dark Web Forum Portal enables access to critical international jihadist and other extremist web forums. Scanlon and Gerber [17] have analyzed messages from the Dark Web portal forums to perform a two-class categorisation task, aimed at distinguishing recruiting posts from non-recruiting posts.

Again, the focus was not on early risk prediction because there was not notion of time or sequence of events.

- **Single actor**. We include in this second category cases where there is not an explicit external actor or intervening factor that causes or stimulates the problem. The risk comes "exclusively" from the individual. For instance, depression might not be caused or stimulated by any intervention or action made by external individuals. Of course, there might be multiple personal or contextual factors that affect –or even cause– a depression process (and, as a matter of fact, this is usually the case). However, it is not feasible to have access to sources of data associated to all these external conditions. In such cases, the only element that can be analysed is the language of the individual. Following this type of analysis, there is literature on the language of people suffering from depression [14, 2, 3, 12], post-traumatic stress disorder [4, 1], bipolar disorder [7], or teenage distress [5]. In a similar vein, other studies have analysed the language of school shooters [18], terrorists [9], and other self-destructive killers [8].

The two classes of risks described above might be related. For instance, individuals suffering from major depression might be more inclined to fall prey to criminal networks. From a technological perspective, different types of tools are likely needed to develop early warning systems that alert about these two types of risks.

Early risk detection technologies can be adopted in a wide range of domains. For instance, it might be used for monitoring different types of activism, studying psychological disorder evolution, early-warning about sociopath outbreaks, or tracking health-related problems in Social Media.

Essentially, we can understand early risk prediction as a process of sequential evidence accumulation where alerts are made when there is enough evidence about a certain type of risk. For the single actor type of risk, the pieces of evidence could come in the form of a chronological sequence of entries written by a tormented subject in Social Media. For the multiple actor type of risk, the pieces of evidence could come in the form of a series of messages interchanged by an offender and a victim in a chatroom or online forum.

To foster discussion on these issues, we shared with the participants of the lab the test collection presented at CLEF in 2016 [10]. This CLEF 2016 paper discusses the creation of a benchmark on depression and language use that formally defines an early risk detection framework and proposes new effectiveness metrics to compare algorithms that address this detection challenge. The framework and evaluation methodology has the potential to be employed by many other research teams across a wide range of areas to evaluate solutions that infer behavioural patterns –and their evolution– in online activity. We therefore invited eRisk participants to engage in a pilot task on early detection of depression, which is described in the next section.

2 Pilot task: Early Detection of Depression

This was an exploratory task on early risk detection of depression. The challenge consists of sequentially processing pieces of evidence and detect early traces of depression

	Train		Test	
	Depressed	*Control*	*Depressed*	*Control*
Num. subjects	83	403	52	349
Num. submissions (posts & comments)	30,851	264,172	18,706	217,665
Avg num. of submissions per subject	371.7	655.5	359.7	623.7
Avg num. of days from first to last submission	572.7	626.6	608.31	623.2
Avg num. words per submission	27.6	21.3	26.9	22.5

Table 1. Main statistics of the train and test collections

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 should 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 pilot task is the collection described in [10]. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, depressed and non-depressed, and, for each user, the collection contains a sequence of writings (in chronological order). For each user, his collection of writings has been divided into 10 chunks. The first chunk contains the oldest 10% of the messages, the second chunk contains 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 a training stage where we released the whole history of writings for a set of training users. We provided all chunks of all training users, and we indicated what users had explicitly mentioned that they have been diagnosed with depression. The participants could therefore tune their systems with the training data. This training dataset was released on Nov 30th, 2016.
- **Test stage**. The test stage consisted of 10 sequential releases of data (done at different dates). The first release consisted of the 1st chunk of data (oldest writings of all test users), the second release consisted of the 2nd chunk of data (second oldest writings of all test users), and so forth. After each release, the participants had one week to process the data and, before the next release, each participating system had to choose between two options: 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 collection. If the system emitted a decision then its decision was considered as final. The systems were evaluated based on the correctness of the decisions and the number of chunks required to make the decisions (see below). The first release was done on Feb 2nd, 2017 and the last (10th) release was done on April 10th, 2017.

Table 1 reports the main statistics of the train and test collections. Both collections are unbalanced (more non-depression cases than depression cases). The number of subjects is not very high, but each subject has a long history of writings (on average, we have hundreds of messages from each subject). Furthermore, the mean range of dates from the first to the last submission is quite wide (more than 500 days). Such wide chronology permits to study the evolution of the language from the oldest piece of evidence to the most recent one.

2.1 Error measure

We employed ERDE, an error measure for early risk detection defined in [10]. This was an exploratory task and the evaluation was tentative. As a matter of fact, one of the goals of eRisk 2017 was to identify the shortcomings of the collection and error metric.

ERDE is a metric for which the fewer writings required to make the alert, the better. For each user we proceed as follows. Given a chunk of data, if a system does not emit a decision then it has access to the next chunk of data (i.e. more writings from the same user). But the system gets a penalty for *late emission*).

Standard classification measures, such as the F-measure, could be employed to assess the system's output with respect to golden truth judgments that inform us about what subjects are really positive cases. However, standard classification measures are time-unaware and, therefore, we needed to complement them with new measures that reward early alerts.

ERDE stands for *early risk detection error* and it takes into account the correctness of the (binary) decision and the delay taken by the system to make the decision. The delay was measured by counting the number (k) of distinct textual items seen before giving the answer. For example, imagine a user u that has 25 writings in each chunk. If a system emitted a decision for user u after the second chunk of data then the delay k was set to 50 (because the system needed to see 50 pieces of evidence in order to make its decision).

Another important factor is that, in many application domains, data are unbalanced (many more negative cases than positive cases). This was also the case in our data (many more non-depressed individuals). Hence, we also needed to weight different errors in a different way.

Consider a binary decision d taken by a 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 face detection tasks where the number of negative cases is several orders of magnitude greater than the number of positive cases. Hence, if we want to avoid building trivial classifiers that always say no, we need to have $c_{fn} >> c_{fp}$. We fixed c_{fn} to 1 and set c_{fp} according to the proportion of positive cases in the test data (e.g. we set c_{fp} to 0.1296). The factor $lc_o(k)(\in [0,1])$ encodes 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 (i.e. late detection is equivalent to not detecting the case at all).

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 (Figure 1 plots $lc_5(k)$ and $lc_{50}(k)$).

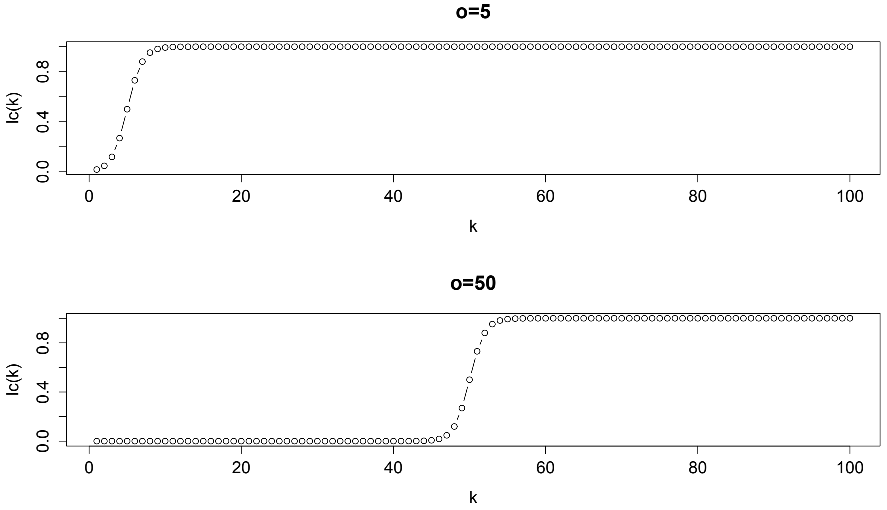

Fig. 1. Latency Cost Functions: $lc_5(k)$ and $lc_{50}(k)$

Observe that the latency cost factor was introduced only for the true positives. We understand that late detection is not an issue for true negatives. True negatives are non-risk cases that, in practice, would not demand early intervention. They just need to be effectively filtered out from the positive cases. Algorithms should therefore focus on early detecting risk cases and detecting non-risk cases (regardless of when these non-risk cases are detected).

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 Contributing systems

We received 30 contributions from 8 different institutions. Table 2 shows the institutions that contributed to eRisk and the labels associated to their runs. Each team could contribute up to five different variants.

Institution	Submitted files
ENSEEIHT, France	GPLA
	GPLB
	GPLC
	GPLD
FH Dortmund, Germany	FHDOA
	FHDOB
	FHDOC
	FHDOD
	FHDOE
U. Arizona, USA	UArizonaA
	UArizonaB
	UArizonaC
	UArizonaD
	UArizonaE
U. Autónoma Metropolitana, Mexico	LyRA
	LyRB
	LyRC
	LyRD
	LyRE
U. Nacional de San Luis, Argentina	UNSLA
U. of Quebec in Montreal, Canada	UQAMA
	UQAMB
	UQAMC
	UQAMD
	UQAME
UACH-INAOE, Mexico-USA	CHEPEA
	CHEPEB
	CHEPEC
	CHEPED
ISA FRCCSC RAS, Russia	NLPISA

Table 2. Participating institutions and submitted results

Next, we briefly describe the main characteristics of the early detection systems implemented by these participants:

- ENSEEIHT, France. This is a joint collaboration between multiple French institutions. They followed a machine learning approach that relies on various features. The best combination was obtained with lexical and statistical features. The user posts were represented with lexicon-based (derived from –or inspired by– previous studies) and numerical features. For example, emotion words (from WordNet Affect) and sentiment words (from Vader) are two of the lexicon-based features defined in the paper. Seven numerical features were defined. For example, the average number of posts or the average number of words per post.
- FH Dortmund, Germany. The Biomedical Computer Science Group from the University of Applied Sciences and Arts Dortmund (FHDO) submitted results obtained from five different models. These models employ linguistic meta information extracted from the users' texts. This team considered classifiers based on Bag of Words, Paragraph Vector, Latent Semantic Analysis (LSA), and Recurrent Neural Networks using Long Short Term Memory (LSTM).
  First, the authors conducted an initial exploratory analysis of the training collection and, next, the authors defined multiple strategies to build their estimations for the test collection. This team has considered a wide range of features: readability features, LIWC features [16, 15] (e.g., statistics on use of pronouns or verb tense), hand-crafted features (for example, specific terms related to antidepressants or diagnosis), and has put in practice sophisticated approaches based on LSTM, neural networks, LSA, and vectorial representation of texts.
- U. Arizona, USA. This team, from the School of Information of the University of Arizona, leveraged external information beyond the available training set. This included a preexisting depression lexicon and concepts extracted from the Unified Medical Language System (UMLS). With these features, they employed sequential –recurrent neural networks– and non-sequential –support vector machines– models for prediction. They also used ensemble methods to leverage the best of each model.
- U. Autónoma Metropolitana, Mexico. The Language and Reasoning Research Group, from Universidad Autónoma Metropolitana (UAM) in México, addressed the challenge with graph models. A graph-based representation was used to capture some inherent characteristics of the documents (computing traditional graph measures). Measurements derived from such representations were then employed as features in a k-nearest neighbour classification system. Contrary to standard Bag of Words representations, these graph models can handle the order of the terms' occurrences in the original text. This permits to incorporate valuable contextual information. The graph models were built by associating neighbouring pairs of tokens with edges that denote their frequency of occurrence. Given the training data, this team's approach was based on building a graph for every document and, next, the graphs from documents of the same class are merged. Ideally, the resulting prototype graph captures common patterns in the content and style of the class (e.g., recurring and neighbouring word sequences). At training time, the graph-based features were obtained by computing different similarity metrics between documents and the prototype graphs.

- U. Nacional de San Luis, Argentina. The LIDIC Research Group, from Universidad Nacional de San Luis, submitted a single run to eRisk. It was based on a semantic representation of documents that explicitly considered the partial information that is available in different chunks of data. Such temporal approach was complemented with standard categorization technology. This team analyzed the temporal variation of terms in the provided chunks. At training time, they found that this analysis of temporal variation had some weaknesses and, thus, they opted for combining their methods with standard predictive tools. The contributed run makes predictions based on their own temporal models and other sources of opinion. They considered multiple document representations (Bag of Words, Concise Semantic Analysis, Character 3-grams, LIWC features) and different learning algorithms, including Random Forests, Naive Bayes, and decision trees.

- U. of Quebec in Montreal, Canada. The University of Quebec contributed to eRisk with five variants of their early prediction system. Several supervised learning approaches and information retrieval models were used to estimate the risk of depression. Among the five systems evaluated, the experiments show that combining information retrieval and machine learning approaches gave the best performance. More specifically, the predictive system was based on an ensemble classification approach that combines supervised learning, information retrieval and feature selection. They used different external resources to build the system. For example, depression-related dictionaries and open-source software (Weka and Solr). The information retrieval component of the system considered the test document (text from the test users) as a query, and searched for similar documents. Two search engines were built using two indexes created from the training set. These indexes follow different pre-processing steps. The supervised learning component of the system combined predictions of multiple classification models with different configurations. They considered different feature types (n-grams, dictionary words, Part-of-Speech and user posting frequency) and three classification algorithms (logistic model tree, an ensemble of sequential minimal optimization classifiers, and an ensemble of random forests). A decision algorithm merged the predictions from the information retrieval component and the supervised learning component.

- UACH-INAOE, Mexico-USA. This is a joint collaboration between the Universidad Autónoma de Chihuahua (Mexico), the Instituto Nacional de Astrofísica, Optica y Electrónica (Mexico), and the University of Houston (USA). Their proposal was based on a two-step classification procedure. First, they looked at post level and created basic features. Next, these features were applied at user level to build a profile for each user. The post-level analysis used representations such as unigrams, bigrams and trigrams, together with extra attributes: the hour of the post and a binary attribute (post on a weekend:yes/no). The user level analysis considered some statistics derived from a post-level classification stage. More specifically, the user representation was based on the sequence of predicted labels for each user post (considering different feature spaces). A Naive Bayes classifier was used for post and user-level classification.

2.3 Performance Results

First, let us analyze the behaviour of the algorithms in terms of how quick they were to emit their decisions. Figures 2 and 3 show a boxplot graph of the number of chunks required to make the decisions. The test collection has 401 subjects and, thus, the boxplot associated to each run represents the statistics of 401 cases. Seven variants waited until the last chunk in order to make the decision for all subjects (i.e. no single decision was done before the last chunk). This happened with CHEPEA, CHEPEB, CHEPEC, CHEPED, FHDOE, GPLD, and NLPISA. These seven runs were extremely conservative: they waited to see the whole history of writings for all the individuals and, next, they emitted their decisions (all teams were forced to emit a decision for each user after the last chunk). Many other runs –e.g., UNLSA, LyRA, LyRB, LyRC, LyRD, UArizonaA, UArizonaB, UArizonaE, UQAMA, UQAMB, UQAMC, UQAMD, and UQAME– also took most of the decisions after the last chunk. For example, with UNSLA, 316 out of 401 test subjects had a decision assigned after the 10th chunk. Only a few runs were really quick at emitting decisions. Notably, FHDOC had a median of 3 chunks needed to emit a decision.

Figure 4 and 5 represent a boxplot of the number of writings required by each algorithm in order to emit the decisions. Most of the variants waited to see hundreds of writings for each user. Only a few runs (UArizonaC, FHDOC and FHDOD) had a median number of writings analyzed below 100. This was the first year of the task and it appears that most of the teams have concentrated on the effectiveness of their decisions (rather than on the tradeoff between accuracy and delay). The number of writings per subject has a high variance. Some subjects have only 10 or 20 writings, while other subjects have thousands of writings. In the future, it will be interesting to study the impact of the number of writings on effectiveness. Such study could help to answer questions like: was the availability of more textual data beneficial?. Note that the writings were obtained from a wide range of sources (multiple subcommunities from the same Social Network). So, we wonder how well the algorithms perform when a specific user had many offtopic writings.

A subject whose main (or only) topic of conversation is depression is arguably easier to classify. But the collection contains non-depressed individuals that are active on depression subcommunities. For example, a person that has a close relative suffering from depression. We think that these cases could be false positives in most of the predictions done by the systems. But this hypothesis needs to be validated through further investigations. We will process the system's outputs and analyze the false positives to shed light on this issue.

Figure 6 helps to analyze another aspect of the decisions of the systems. For each group of subjects, it plots the percentage of correct decisions against the number of subjects. For example, the rightmost bar of the upper plot means that 90% of the systems correctly identified one subject as depressed. Similarly, the rightmost bar of the lower plot means that there were 46 non-depressed subjects that were correctly classified by all systems (100% correct decisions). The graphs show that systems tend to be more effective with non-depressed subjects. The distribution of correct decisions for non-depressed subjects has many cases where more than 80% of the systems are correct. The distribution of correct decisions for depressed subjects is flatter, and many depressed subjects are only identified by a low percentage of the systems. Furthermore, there are not depressed subjects that are correctly identified by all systems. However, an interesting point is that no depressed subject has 0% of correct decisions. This means that every depressed subject was classified as such by at least one system.

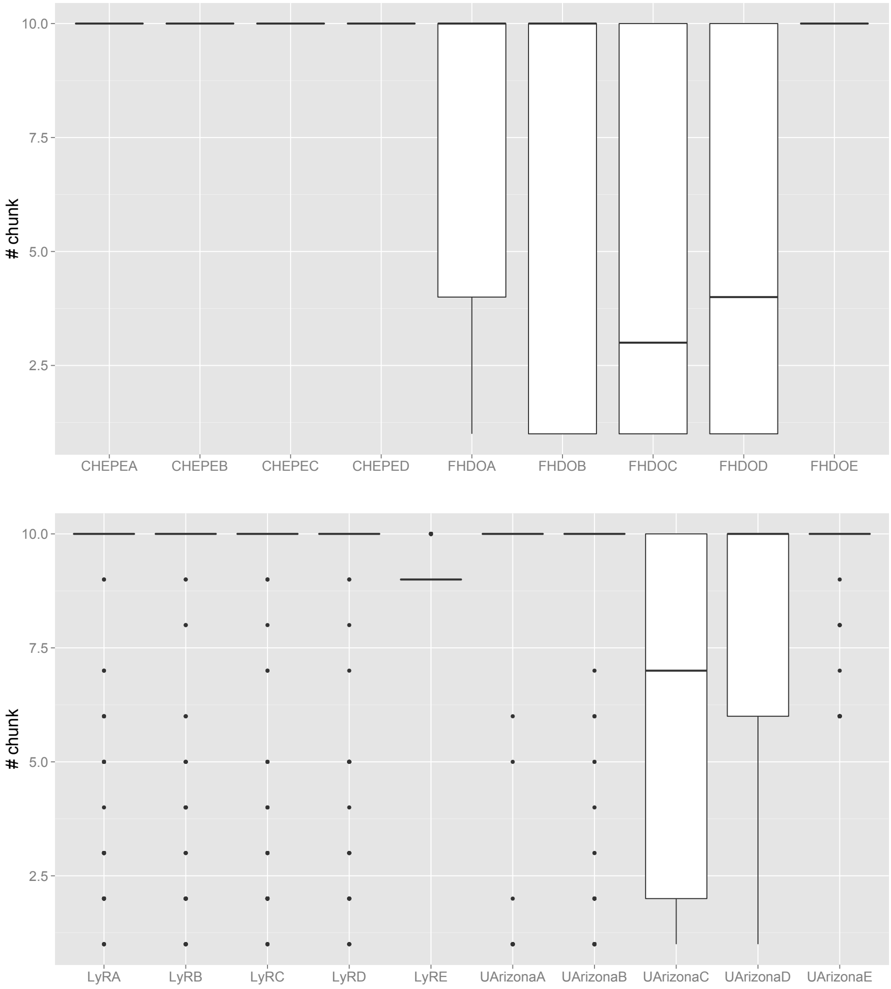

Fig. 2. Number of chunks required by each contributing run in order to emit a decision.

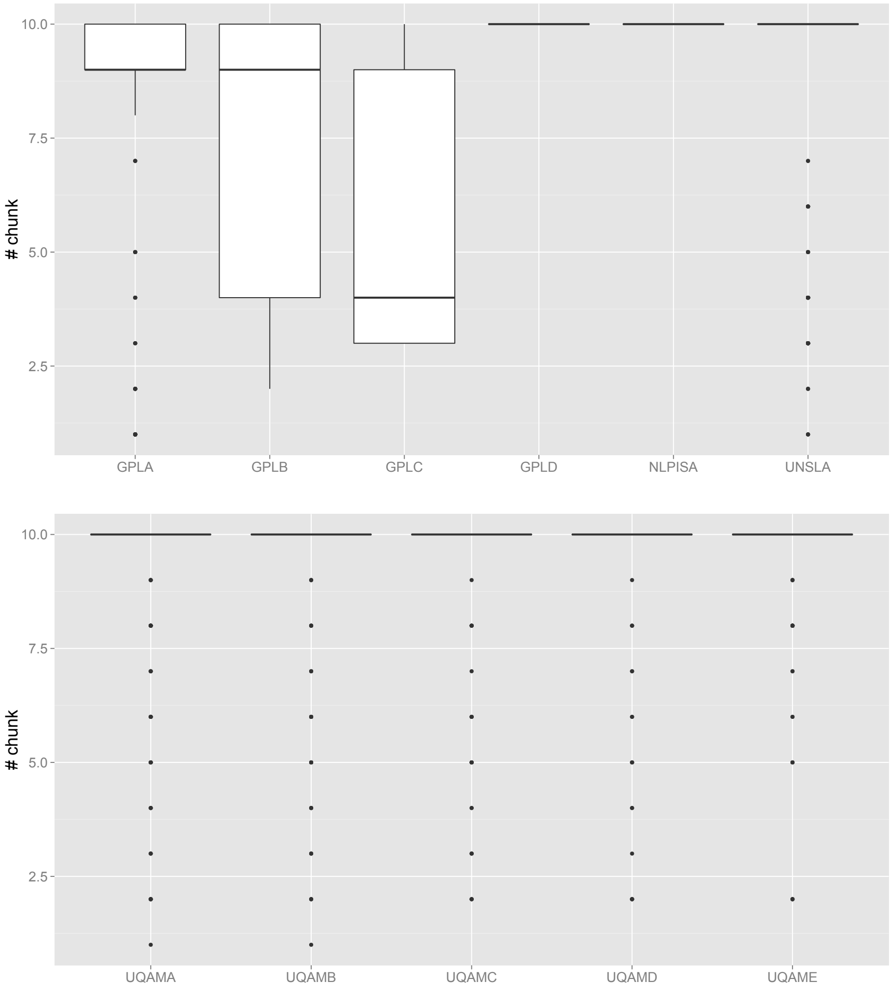

Fig. 3. Number of chunks required by each contributing run in order to emit a decision.

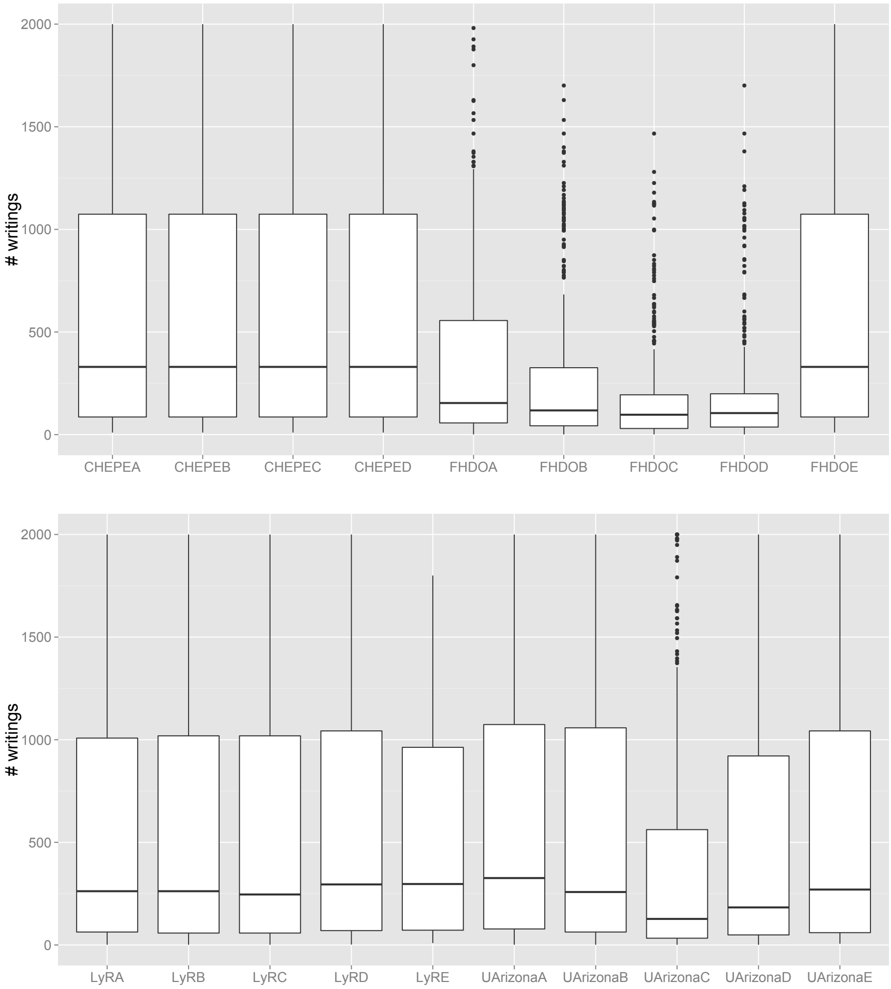

Fig. 4. Number of writings required by each contributing run in order to emit a decision.

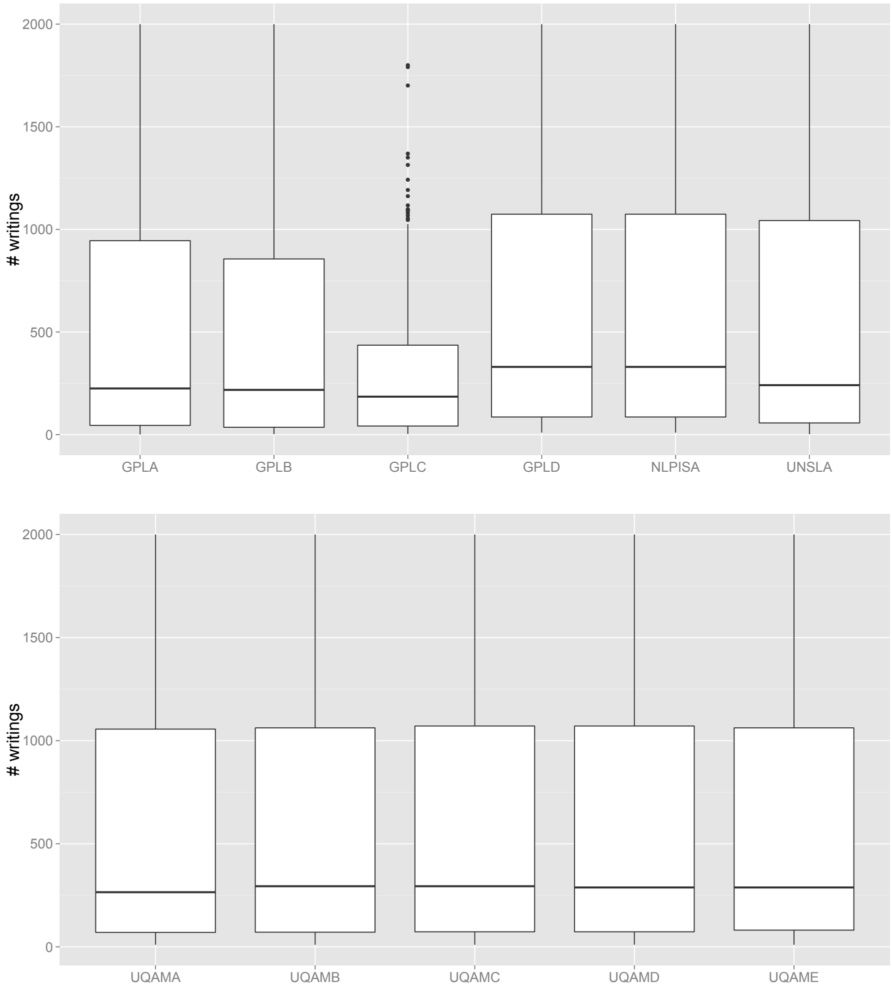

Fig. 5. Number of writings required by each contributing run in order to emit a decision.

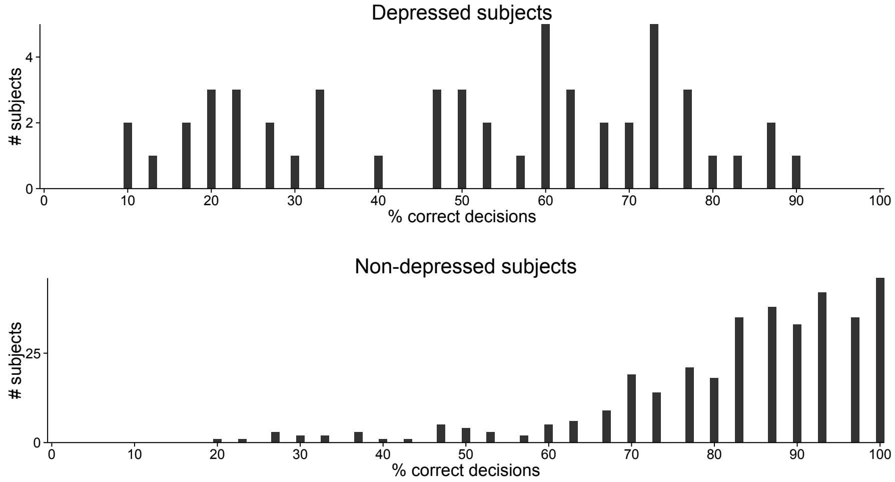

Fig. 6. Number of depressed and non-depressed subjects that had a given percentage of correct decisions.

Let us now analyze the effectiveness results (see Table 3). The first conclusion we can draw is that the task is difficult. In terms of F1, performance is low. The highest F1 is 0.64. This might be related to the way in which the collection was created. The non-depressed group of subjects includes random users of the social networking site, but also a number of users who were active on the depression community and depression fora. There is a variety of such cases but most of them are individuals interested in depression because they have a close relative suffering from depression. These cases could potentially be false positives. As a matter of fact, the highest precision, 0.69, is also relatively low. The lowest $ERDE_5$ was achieved by the FHDO team, which also submitted the runs that performed the best in terms of $F1$ and precision. The run with the lowest $ERDE_{50}$ was submitted by the UNSLA team.

They're binary. computed on positive class

Some systems, e.g. FHDOB, opted for optimizing precision, while other systems, e.g. UArizonaC, opted for optimizing recall. The lowest error tends to be associated with runs with moderate F1 but high precision. For example, FHDOB, the run with the lowest $ERDE_5$, is one of the runs that was quicker at making decisions (see Figs 2 and 4) and its precision is the highest (0.69). $ERDE_5$ 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 $ERDE_{50}$, UNSLA, has low precision but relatively high recall (0.79). Such difference between $ERDE_5$ and $ERDE_{50}$ is highly relevant in practice. For example, a mental health agency seeking a tool for automatic screening for depression could set the penalty costs depending on the consequences of late detection of depression.

	$ERDE_5$	$ERDE_{50}$	F1	P	R
GPLA	17.33%	15.83%	0.35	0.22	0.75
GPLB	19.14%	17.15%	0.30	0.18	0.83
GPLC	14.06%	12.14%	0.46	0.42	0.50
GPLD	14.52%	12.78%	0.47	0.39	0.60
FHDOA	12.82%	9.69%	**0.64**	0.61	0.67
FHDOB	**12.70%**	10.39%	0.55	**0.69**	0.46
FHDOC	13.24%	10.56%	0.56	0.57	0.56
FHDOD	13.04%	10.53%	0.57	0.63	0.52
FHDOE	14.16%	12.42%	0.60	0.51	0.73
UArizonaA	14.62%	12.68%	0.40	0.31	0.58
UArizonaB	13.07%	11.63%	0.30	0.33	0.27
UArizonaC	17.93%	12.74%	0.34	0.21	**0.92**
UArizonaD	14.73%	10.23%	0.45	0.32	0.79
UArizonaE	14.93%	12.01%	0.45	0.34	0.63
LyRA	15.65%	15.15%	0.14	0.11	0.19
LyRB	16.75%	15.76%	0.16	0.11	0.29
LyRC	16.14%	15.51%	0.16	0.12	0.25
LyRD	14.97%	14.47%	0.15	0.13	0.17
LyRE	13.74%	13.74%	0.08	0.11	0.06
UNSLA	13.66%	**9.68%**	0.59	0.48	0.79
UQAMA	14.03%	12.29%	0.53	0.48	0.60
UQAMB	13.78%	12.78%	0.48	0.49	0.46
UQAMC	13.58%	12.83%	0.42	0.50	0.37
UQAMD	13.23%	11.98%	0.38	0.64	0.27
UQAME	13.68%	12.68%	0.39	0.45	0.35
CHEPEA	14.75%	12.26%	0.48	0.38	0.65
CHEPEB	14.78%	12.29%	0.47	0.37	0.63
CHEPEC	14.81%	12.57%	0.46	0.37	0.63
CHEPED	14.81%	12.57%	0.45	0.36	0.62
NLPISA	15.59%	15.59%	0.15	0.12	0.21

Table 3. Results

3 Workshop

The lab was also 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. Potential participants could cover a wide range of areas in information access and closely related fields, such as Natural Language Processing, Machine Learning, and Information Retrieval.

We accepted one paper on Temporal Variation of Terms, submitted jointly by the LIDIC research team of the Universidad Nacional de San Luis and CONICET (Argentina). This research group participated into the pilot task described above, and their submission to the workshop provides a detailed description of a new document representation, named Temporal Variation of Terms (TVT), that was employed in their early detection models.

TVT is a sophisticated approach that takes ideas from previous concept space representations and is based on using the variation of vocabulary along different time steps as a concept space to represent the documents. The method builds on Concise Semantic Analysis (CSA) techniques, and instantiates CSA with concepts derived from terms occurring in the temporal chunks, which are analyzed at different time steps. CSA represents terms in a space of concepts that is equal or close to the category labels; and documents are the centroid of term vectors. The main idea of TVT is to use the temporal information to obtain an extended concept space. The paper analyzes the results obtained with multiple classifiers that work with the eRisk 2017 pilot task data.

4 Conclusions

This paper provides an overview of eRisk 2017. This was the first year that this lab was organized at CLEF and the lab's activities were concentrated on a pilot task on early risk detection of depression. The task received 30 contributions from 8 different institutions. Being the first year of the task, most teams focused on tuning different classification solutions (depressed vs non-depressed). The tradeoff between early detection and accuracy was not a major concern for most of the participants.

Besides the papers submitted by the teams contributing to the pilot task, eRisk 2017 also had a *workshop session*, which included a paper on a new concept space for early risk prediction.

We plan to run eRisk again in 2018. We are currently collecting more data on depression and language, and we plan to expand the lab to other psychological problems. Early detecting other disorders, such as anorexia or post-traumatic stress disorder, would also be highly valuable and could be the focus of some eRisk 2018 subtasks.

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 investigación de Galicia") and ED431G/08.

References

1. Jennifer Alvarez-Conrad, Lori A. Zoellner, and Edna B. Foa. Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, 15(7):S159–S170, 2001.
2. Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In Hugh C. Davis, Harry Halpin, Alex Pentland, Mark Bernstein, and Lada A. Adamic, editors, *WebSci*, pages 47–56. ACM, 2013.

3. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.
4. Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014.
5. Karthik Dinakar, Emily Weinstein, Henry Lieberman, and Robert Louis Selman. Stacked generalization learning to analyze teenage distress. In Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice Oh, editors, *ICWSM*. The AAAI Press, 2014.
6. Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at PAN-2012. In *Proceedings of the PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse (within CLEF 2012)*, 2012.
7. Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. Text analysis as a tool for analyzing conversation in online support groups. In Elizabeth Dykstra-Erickson and Manfred Tscheligi, editors, *CHI Extended Abstracts*, pages 1485–1488. ACM, 2004.
8. Adam Lankford. Précis of the myth of martyrdom: What really drives suicide bombers, rampage shooters, and other self-destructive killers. *Behavioral and Brain Sciences*, 37:351–362, 8 2014.
9. Cecilia H Leonard, George D Annas, James L Knoll, and Terje Tørrissen. The case of Anders Behring Breivik - language of a lone terrorist. *Behav Sci Law*, 2014.
10. 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.
11. India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. Learning to identify internet sexual predation. *Int. J. Electron. Commerce*, 15(3):103–122, April 2011.
12. Megan A. Moreno, Lauren A. Jelenchick, Katie G. Egan, Elizabeth Cox, Henry Young, Kerry E. Gannon, and Tara Becker. Feeling bad on facebook: depression disclosures by college students on a social networking site., June 2011.
13. Javier Parapar, David E. Losada, and Alvaro Barreiro. A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs. In *PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse, at Conference and Labs of the Evaluation Forum CLEF*, Rome, Italy, 2012.
14. Minsu Park, David W. McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in Twitter. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.
15. James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The development and psychometric properties of LIWC2007 @ONLINE, June 2012.
16. J.W. Pennebaker, M.R. Mehl, and K.G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.
17. JacobR Scanlon and MatthewS Gerber. Automatic detection of cyber-recruitment by violent extremists. *Security Informatics*, 3(1), 2014.
18. Jari Veijalainen, Alexander Semenov, and Jorma Kyppö. Tracing potential school shooters in the digital sphere. In Samir Kumar Bandyopadhyay, Wael Adi, Tai-Hoon Kim, and Yang Xiao, editors, *ISA*, volume 76 of *Communications in Computer and Information Science*, pages 163–178. Springer, 2010.
19. Yulei Zhang, Shuo Zeng, Chunneng Huang, Li Fan, Ximing Yu, Yan Dang, Catherine A. Larson, Dorothy Denning, Nancy Roberts, and Hsinchun Chen. Developing a dark web collection and infrastructure for computational and social sciences. In Christopher C. Yang, Daniel Zeng, Ke Wang, Antonio Sanfilippo, Herbert H. Tsang, Min-Yuh Day, Uwe Glässer, Patricia L. Brantingham, and Hsinchun Chen, editors, *ISI*, pages 59–64. IEEE, 2010.