UNSL's participation at eRisk 2018 Lab

Dario G. Funez[1], Ma. José Garciarena Ucelay[1], Ma. Paula Villegas[1], Sergio G. Burdisso[1,2], Leticia C. Cagnina[1,2], Manuel Montes-y-Gómez[3], and Marcelo L. Errecalde[1]

[1] LIDIC Research Group, Universidad Nacional de San Luis, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
[3] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
{funezdario,mjgarciarenaucelay,villegasmariapaula74}@gmail.com,
{sergio.burdisso,lcagnina}@gmail.com, mmontesg@inaoep.mx,
merrecalde@gmail.com

Abstract. In this paper we describe the participation of the LIDIC Research Group of Universidad Nacional de San Luis (UNSL) - Argentina at CLEF eRisk 2018 Lab. The main goal of this Lab is considering early risk detection scenarios where the issue of getting timely predictions with a reasonable confidence level becomes critical. Two completely different approaches were used, that we will refer as *flexible temporal variation of terms* (FTVT) and *sequential incremental classification* (SIC). FTVT is a semantic representation of documents that explicitly considers the partial information that is made available in the different "chunks" to the early risk detection systems along the time. FTVT is an improvement on the TVT method [1] that allows varying the number of chunks considered in the representation according to the "level of urgency" required in the classification. SIC is a novel approach for text categorization that incrementally estimates the level of belonging of a piece of text to the different categories based on an accumulative process of evidence. In the test stage, FTVT obtained the lowest $ERDE_5$ error in both pilot tasks and SIC achieved the highest precision for the anorexia detection task providing strong evidence that both approaches used by our team are interesting alternatives to deal with early risk detection tasks.
Keywords: Early Risk Detection, Early Depression Detection, Early Anorexia Detection, Semantic Analysis Techniques, Flexible Temporal Variation of Terms, Incremental Classification.

1 Introduction

The increasing use of Internet, social networks and other computer technologies allows the extraction of valuable information to early prevent some risks. In this context, early risk detection (ERD) on the Internet is an important research area due to the impact it might have in areas like *health* when people suffer depression, anorexia or other disorders that can threaten life and *safety* when criminals and sex offenders try to attack using web technologies.

The same as other predictive tasks, ERD methods have been mainly based on supervised machine learning approaches. In those cases, the task is generally addressed as a standard binary classification problem with two unbalanced classes: a minority

(risky) positive class and a majority (control) negative class. However, beyond the difficulty that the unbalanced classes present to the learning algorithms, ERD introduces an added problem that is not usually present in other classification tasks: the incremental classification of sequential data (ICSD).

To effectively support ICSD two important aspects need to be considered. First, we must provide an adequate way to "remember" or "summarize" historical information read up to specific points of time. The informativeness level of these partial models will be critical to the effectiveness of the classifier in charge of detecting risky cases. Second, these models need to also provide support to a very important aspect of ERD: the decision of *when* (how soon) the system should stop reading from the input stream and classify it with an acceptable level of accuracy. This aspect, that we will refer as the *supporting for early classification*, is basically a multi-objective decision problem that attempts balancing accurate and timely classifications [7]. In fact, common evaluation measures of supervised classification like precision, recall and F-measure are no longer adequate in those cases because they do not take "time" into account. Thus, new "temporal" measures that penalize the system's delay in detecting risky cases are required. This is the case of the $ERDE_o$ error introduced in [4] and used in the 2017 eRisk pilot task [5] which allows specifying a threshold (the o value) that, when is surpassed, the penalty rapidly grows to 1.

The eRisk 2018 Lab presented two challenging tasks for ERD: early detection of signs of depression (task 1), and early detection of signs of anorexia (task 2). We participated in both tasks with two different approaches to deal with the ICSD issue: one that we will refer as *flexible temporal variation of terms* (FTVT) and the other named *sequential incremental classification* (SIC).

FTVT is a document representation that deals with the ICSD problem by keeping sequential information about the variation of terms occurring in the different chunks. The hypothesis behind this approach is that these variations can be informative to detect a risky case. SIC is a sequential approach that incrementally reads and estimates the evidence that words provide for both, positive and negative classes. SIC classifies a subject as risky as soon as the accumulated evidence of the risky (positive) class surpass the evidence of the negative one.

The experiments carried out on the training sets for both tasks were mainly aimed at determining adequate parameters for training the models (classifiers) for the test stage. Preliminary results reported by the Lab's organizers showed that our systems obtained the best (lowest) $ERDE_5$ error in both pilot tasks and SIC the highest precision for the anorexia detection task providing strong evidence that the used approaches are interesting alternatives to deal with early risk detection tasks.

The rest of the article is organized as follows: Section 2 gives general information of the data sets used in both pilot tasks and the methods used in our ERD systems. Next, in Section 3 the activities carried out in the training stage are described and the rationale behind the main design decisions made on our ERD systems, are presented. Section 4 shows the performance of our methods on the eRisk 2018 data sets released in the test stage. Finally, Section 5 depicts potential future works and the obtained conclusions.

2 Data sets and methods

2.1 Data Sets

The data sets supplied for the eRisk 2018 tasks[4] are described in Losada et al. [6]. Both collections (for task 1 and task 2) of writings (post or comments) were extracted from Social Media. For each user (document in the data sets), the collections contain a sequence of writings (in chronological order) which has been partitioned 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 corpus of task 1 is related to depression and for the task 2 is about anorexia. In the first one, there are two categories of users, "depressed" and "non-depressed", meanwhile in the corpus of anorexia the users are "anorexic" and "non-anorexic". The collection of depression was split into a training and a test set that we will refer as $\mathcal{TR}_{DS}$ and $\mathcal{TE}_{DS}$, respectively. The $\mathcal{TR}_{DS}$ set contains 887 users (135 positive, 752 negative) and the $\mathcal{TE}_{DS}$ set contains 820 users (79 positive, 741 negative). The users labeled as positive are those that have explicitly mentioned that they have been diagnosed with depression. The corpus of anorexia was split into a training and a test set that we will refer as $\mathcal{TR}_{AX}$ and $\mathcal{TE}_{AX}$, respectively. The $\mathcal{TR}_{AX}$ set contains 152 users (20 positive, 132 negative) and the $\mathcal{TE}_{AX}$ set contains 320 users (41 positive, 279 negative). In this case, the users labeled as positive are those that have been diagnosed with anorexia.
Each task was divided by their organizers into a *training stage* and a *test stage*. In the first one, the participating teams had access to the set of training users with ten chunks of all training users. They could therefore tune their systems with the training data. Then, in the *test stage*, the ten chunks from test set were gradually released by the organizers one by one until completing all the chunks that correspond to the complete writings of the considered individuals. Each time that a chunk ch_i was released, participants in the pilot tasks were asked to give their predictions on the users contained in the test set, based on the partial information read from chunks ch_1 to ch_i. Once a class of an incoming stream is predicted, that decision is irreversible (it cannot be undone).

2.2 Methods

To deal with the problems posed in both pilot tasks we used two methods that previously referred as FTVT and HCI, which will describe below. An interesting aspect of those methods is that they are completely *independent-domain*. Thus, they do not require costly adaptation processes for each task, beyond the tuning of parameters that could depend of the used data set. In fact, due to limitations of time to carry out the experimental study, both methods were only evaluated on the data set of the task 1 (early depression detection) and the same parameters were used for task 2 (early anorexia detection).

Space constraints prevent us from giving detailed explanations of FTVT and HCI. However, the interested reader can obtain in [1] more implementation details of the

[4] http://early.irlab.org/task.html

TVT method on which FTVT is based on. SIC is only introduced from an intuitive point of view because the method is currently under review in a scientific journal.[5]

Flexible Temporal Variation of Terms (FTVT) The *Flexible Temporal Variation of Terms* (FTVT) is an improvement of the *temporal variation of terms* (TVT) method [1], an approach for early risk detection that uses the temporal variation of terms between chunks as concept space of a concise semantic analysis (CSA) approach [2]. The main characteristic of the original TVT is that it allowed to address the unbalance of the minority class with information of the first 4 "chunks" of the users (that number was determined empirically). FTVT provides a more flexible approach than TVT by allowing the specification of a different number of chunks n for the distinct systems. This small extension on TVT is not a minor aspect. Several studies with FTVT showed that, depending on the urgency level required for the ERD task (determined by the threshold o) the number n used in FTVT produces very different $ERDE_o$ values. However, beyond this small difference between TVT and FTVT, there is no conceptual differences between both approaches and, therefore, we will only give a short description of the original TVT approach.

As we previously said, TVT is based on the concise semantic analysis (CSA) technique proposed in [2] and later extended in [3] for author profiling tasks. CSA is a semantic analysis technique that interprets words and text fragments in a space of concepts that are close (or equal) to the category labels. For instance, if documents in the data set are labeled with q different category labels (usually no more than 100 elements), words and documents will be represented in a q-dimensional space. That space size is usually much smaller than standard BoW representations which directly depend on the vocabulary size (more than 10000 or 20000 elements in general). CSA has been used in general text categorization tasks [2] and has been adapted to work in author profiling tasks under the name of Second Order Attributes (SOA) [3].

In this context, the underlying idea of TVT is that variations of the terms used in different sequential stages of the documents may have relevant information for the classification task. With this idea in mind, this method enriches the documents of the minority class with the partial documents read in the first 4 chunks. These chunks correspond to the minority (depressed or positive) class. Also TVT uses the complete documents (chunk 10). All this information is considered as a new concept space for a CSA method.

TVT naturally copes with the sequential caracteristics of ERD problems and also gives a tool for dealing with unbalanced data sets. Preliminary results of this method in comparison to CSA and BoW representations [1] showed its potential to deal with ERD problems. FTVT, the variant of TVT used in the present work, arose from our observation that, varying the number n of initial chunks, different performance can be achieved depending on the $ERDE_o$ measure used to evaluate the results.

[5] The person interested in deeper technical details of both methods can obtain more information in https://sites.google.com/site/lcagnina/technicalreport-ftvt and https://sites.google.com/site/lcagnina/technicalreport-sic.

Sequential Incremental Classification (SIC) Sequential Incremental Classification (SIC) is a very simple method. During the training phase a dictionary of words is built for each category, in which frequency of each word is stored. Then, using those word frequencies, and during classification stage, a value for each word was calculated using a function $gv(w, c)$ to value words in relation to categories. gv takes a word w and a category c and outputs a number in the interval [0,1] representing the degree of confidence with which w is believed to *exclusively* belong to c, for instance, suppose categories $C = \{food, music, health, sports\}$, we could have:

$$gv(\text{`}sushi\text{'}, food) = 0.85; \quad gv(\text{`}the\text{'}, food) = 0;$$
$$gv(\text{`}sushi\text{'}, music) = 0.09; \; gv(\text{`}the\text{'}, music) = 0;$$
$$gv(\text{`}sushi\text{'}, health) = 0.50; \; gv(\text{`}the\text{'}, health) = 0;$$
$$gv(\text{`}sushi\text{'}, sports) = 0.02; \; gv(\text{`}the\text{'}, sports) = 0;$$

Additionally, $\overrightarrow{gv}(w) = (gv(w, c_0), gv(w, c_1), \ldots, gv(w, c_k))$ is defined, where $c_i \in C$ (the set of all the categories). That is, $\overrightarrow{gv}$ is only applied to a word and it outputs a vector in which each component is the *gv* of that word for each category c_i. For instance, following the above example, we have:

$$gv(\text{`}sushi\text{`}) = (0.85, 0.09, 0.5, 0.02); \; gv(\text{`}the\text{`}) = (0, 0, 0, 0);$$

We have called the vector $\overrightarrow{gv}(w)$, the "*confidence vector* of w". Note that each category c_i is assigned a fixed position, i, in $\overrightarrow{gv}$ (for instance, in the example above $(0, 0, 0, 0)$ is the *confidence vector* of "the" and the first position corresponds to $food$, the second to $music$, and so on).

Classification is finally carried out, for each subject, by means of the cumulative sum of all words $\overrightarrow{gv}$ vectors, in symbols:

$$\overrightarrow{d} = \sum_{w \in S} \overrightarrow{gv}(w)$$

where S is the subject's writing history. Note that $\overrightarrow{d}$ is a vector with two components, one for the positive class (depressed or anorexic) and one for the negative (control) class. The policy to classify a subject as positive was performed by analyzing how $\overrightarrow{d}$ changed over time (i.e. over "chunks"), as shown with an example in Figure 1 for a depression case. Subjects were classified as depressed when the cumulated positive value exceeded the negative one, for instance the subject in the figure was classified as depressed after reading the 5th chunk.

It is worth mentioning that, to compute gv we used other two functions, lv and $weight$, as follows:

$$gv(w, c) = lv_\sigma(w, c) \times weight_\lambda(w, c)$$

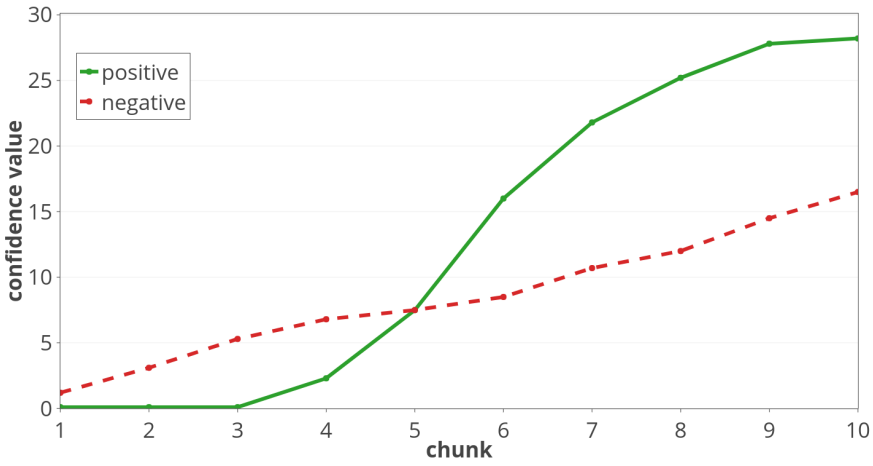

Fig. 1. Subject-9579's cumulated positive and negative confidence values variation over time (chunks).

- $lv_\sigma(w, c)$ values a word based on the local frequency of w in c. As part of this process, the word distribution curve is smoothed by a factor controlled by the hyperparameter σ.
- $weight_\lambda(w, c)$ decreases lv in relation to the lv value of w to the other categories. The more categories c_i whose $lv_\sigma(w, c_i)$ is high, the smaller the $weight_\lambda(w, c)$ value. The λ hyperparameter controls how sensitive this sanction is.

3 Experimental Setting

As we mentioned above, this year there were two tasks: one for the early detection of depression cases (task 1) and the other one for the early detection of people with anorexia (task 2). We only used the $\mathcal{TR}_{DS}$ data set for setting the parameters of our methods because it is the largest and, therefore, it would seem to be more appropriate to obtain more confident statistics.

In order to find the best values for the parameters of our methods we perform a five-fold cross validation on the depression training set. Hence, we divided the $\mathcal{TR}_{DS}$ set into five folds (see Table 1). These folds maintain the same proportions of both kind of users and were randomly selected. Also each fold was divided into 10 chunks like they were provided by the organizers. We trained the classifiers with four folds and tested with the fifth fold. This process was repeated four times more, always choosing different folds, and later the results were averaged.

We used the *Flexible Temporal Variation of Terms* (FTVT) described previously to represent the documents. For this representation, a decision must be made related to the number n of chunks that will enrich the minority (positive) class. We considered

Table 1. Distribution of the $\mathcal{TR}_{DS}$ set in folds.

Fold	Positive	Negative	Total
1	27	150	177
2	27	151	178
3	27	151	178
4	27	150	178
5	27	150	177
Total	135	752	887

different values for n, particularly we selected n from 0 to 5 for setting the initial chunks used.

FTVT was evaluated with different learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB), among others. We used the implementation provided in the Scikit-learn package for Python 2.7 with the default parameters. That is, $penalty = l2$ and $C = 1$, for both SVM and LR.

We used the probability p assigned by the classifier to decide *when* to stop reading a document and giving its classification. Thus, our approach considered that when the probability p assigned to the positive class exceeds some particular threshold θ ($p \geq \theta$) the instance/document is classified as positive. We used different thresholds θ: 0.9, 0.8, 0.7 and 0.6.

We evaluated the performance of our approaches with the *early risk detection error* (ERDE) measure proposed in [4]. This measure takes into account not only the correctness of the decision made by the system but also the delay in making that decision. ERDE uses specific costs to penalize false positives and false negatives. However, ERDE has a different treatment with the two possible successful predictions (true negatives and true positives). True negatives have no cost (cost$= 0$) but ERDE associates a cost to the delay in the detection of true positives that monotonically increases with the number k of textual items seen before giving the answer. In a nutshell, that cost is low when k is lower than a threshold value o but rapidly approaches 1 when $k > o$. In that way, o represents some type of "urgency" in detecting depression cases: the lowest the o values the highest the urgency in detecting the positive cases. A more detailed description of ERDE can be found in [4]. We considered the two values of o employed in both editions (2017 and 2018) of this pilot task: $o = 5$ ($ERDE_5$) and $o = 50$ ($ERDE_{50}$).

Due to space constraints, only combinations of n (parameter of FTVT), θ (probability threshold to classify an instance as positive) and the used classifier that allowed obtaining the best values of $ERDE_5$ and $ERDE_{50}$ metrics, are shown. These results are presented in Tables 2 and 3, respectively.

If we analyze Table 2, it can be seen that small values for n and a high threshold θ (more restrictive) generate a lower $ERDE_5$. In particular, the best configuration for $ERDE_5$ is $n = 0$ and $p \geq 0.8$ with the SVM algorithm obtaining 13.58. A higher threshold means that it is necessary more confidence to classify a user as positive. This is because as the urgency level to decide is also high, what can be classified as positive has to be precise, otherwise the penalty is higher. On the other hand, with regards to the $ERDE_{50}$ metric we can see in Table 3 that the best thresholds are a little lower than in

Table 2. Best performance of FTVT for $ERDE_5$ metric.

n	Classifier	θ	$ERDE_5$
0	**SVM**	**0.8**	**13.58**
1	LR	0.9	13.75
2	LR	0.8	13.74
3	LR	0.9	13.68
4	SVM	0.9	13.84
5	SVM	0.9	13.87

Table 3. Best performance of FTVT for $ERDE_{50}$ metric.

n	Classifier	θ	$ERDE_{50}$
0	SVM	0.6	10.25
1	SVM (or LR)	0.6	9.91
2	LR	0.6	9.61
3	SVM	0.6	9.77
4	**SVM**	**0.7**	**9.59**
5	SVM (or LR)	0.7	9.69

Table 2 ($\theta = 0.6$ and $\theta = 0.7$). In particular, the best result is 9.59 and is obtained with $n = 4$, $p \geq 0.7$ and SVM as classifier. However, the performance achieved when $n = 2$ is also good enough.

From these results we can conclude that FTVT with $n = 0$ in combination with the SVM classifier and a probability threshold $\theta = 0.8$ seems to be adequate for the $ERDE_5$. Hereafter, this configuration will be referred as $UNSLA$. The FTVT with $n = 2$ in combination with the Logistic Regression and $\theta = 0.6$ seems to be an adequate balance between $ERDE_5$ and $ERDE_{50}$ ($UNSLB$). Finally, FTVT with $n = 4$ in combination with SVM and $\theta = 0.7$ looks as a reasonable alternative for the $ERDE_{50}$ metric ($UNSLC$).

Regarding SIC, no hyper-parameter optimization was done and the same hyper-parameter values were used for both tasks (anorexia and depression). Hyperparameters values were arbitrarily set to $\sigma = 0.5$ for both $UNSLD$ and $UNSLE$, and $\lambda = 3$ and $\lambda = 7$ for $UNSLD$ and $UNSLE$, respectively. $UNSLD$ was meant to be less sensitive on penalizing words and thus considering more words as being "important" than $UNSLE$, hence favoring $UNSLD$ to have a higher recall but with the risk of having a worse precision.

In summary, from the above study we selected the settings showed in Table 4 to participate in the 2018 pilot tasks.

4 Evaluation Stage

Our five systems, three variants of FTVT ($UNSLA$, $UNSLB$ and $UNSLC$) and two variants of SIC ($UNSLD$ and $UNSLE$) were trained with the full training set of the pilot task 1 ($\mathcal{TR}_{DS}$) and tested with the corresponding $\mathcal{TE}_{DS}$ (see Table 5).In the same

Table 4. Settings of the submitted approaches.

Submitted approach	Method	Parameters	Learning algorithm
$UNSLA$	FTVT	$n = 0, \theta = 0.8$	SVM
$UNSLB$	FTVT	$n = 2, \theta = 0.6$	LR
$UNSLC$	FTVT	$n = 4, \theta = 0.7$	SVM
$UNSLD$	SIC	$\sigma = 0.5, \lambda = 3$	SIC
$UNSLE$	SIC	$\sigma = 0.5, \lambda = 7$	SIC

way, for task 2 the methods were trained with $\mathcal{TR}_{AX}$ and tested with the corresponding $\mathcal{TE}_{AX}$ (see Table 6). Both test sets were incrementally released during the testing phase of the pilot tasks.

Table 5. Best in the ranking of the *depression* pilot task ($\mathcal{TE}_{DS}$ set).

	$ERDE_5$	$ERDE_{50}$	F_1	π	ρ
$UNSLA$	**8.78**	7.39	0.38	0.48	0.32
$UNSLB$	8.94	7.24	0.40	0.35	0.46
$UNSLC$	8.82	6.95	0.43	0.38	0.49
$UNSLD$	10.68	7.84	0.45	0.31	0.85
$UNSLE$	9.86	7.60	0.60	0.53	0.70
FHDO-BCSGB	9.50	**6.44**	**0.64**	0.64	0.65
RKMVERIC	9.81	9.08	0.48	**0.67**	0.38
UDCB	15.79	11.95	0.18	0.10	**0.95**

In Table 5 we show the results of our 5 submissions and the results of those systems that obtained the best $ERDE_5$, $ERDE_{50}$, F_1, precision and recall in the eRisk *depression* pilot task as reported in [6]. Best values are highlighted in boldface. There, we can observe that our $UNSLA$ obtained the best $ERDE_5$ value. On the other hand, FHDO-BCSGB achieved the best $ERDE_{50}$ and F-measure although our $UNSLC$ obtained a value quite similar (slightly worse) for $ERDE_{50}$. $UNSLE$ obtained the 3th best F_1(0.60) measure (the 1st and the 2nd one belonged to the FHDO-BCSG team) and $UNSLD$ obtained the 2nd best recall (0.85) measure[6].

Table 6 shows similar results for the *anorexia* pilot task. As we can see, our system ($UNSLB$ in this case) obtained the best $ERDE_5$ again and $UNSLD$ the best precision value. At this point it is important to note that we did not perform a parameter optimization of our methods for the anorexia task, such as we stated in the previous section. Then, it is not a minor aspect that our systems can perform well in a different domain from the used for setting the parameters. This independence of domain is such a really important aspect of the classifier systems for the optimization of real tasks.

With these results, we can conclude that our proposals are very reasonable and competitive alternatives for ERD tasks.

[6] Although, the 1st one (UDCB) had a very low precision (0.1).

Table 6. Best in the ranking of the *anorexia* pilot task ($\mathcal{TE}_{AX}$ set).

	$ERDE_5$	$ERDE_{50}$	F_1	π	ρ
$UNSLA$	12.48	12.00	0.17	0.57	0.10
$UNSLB$	**11.40**	7.82	0.61	0.75	0.51
$UNSLC$	11.61	7.82	0.61	0.75	0.51
$UNSLD$	12.93	9.85	0.79	**0.91**	0.71
$UNSLE$	12.93	10.13	0.74	0.90	0.63
FHDO-BCSGD	12.15	**5.96**	0.81	0.75	**0.88**
FHDO-BCSGE	11.98	6.61	**0.85**	0.87	0.83

5 Conclusions and future work

This article presented the participation of UNSL at eRisk 2018 Pilot tasks on Early Detection of Depression and Anorexia. We used two completely different approaches to deal with those tasks: one based on the FTVT representation and other on a simple method named SIC. Those approaches showed to be very effective on both types of tasks obtaining the best $ERDE_5$ value over all participants in both tasks and the best precision value for the anorexia task. Besides, in the $ERDE_{50}$ measure, although we did not achieve the best value, our results were very close to it. Thus, the performance of our systems seem to indicate that the used methods are very robust approaches for ERD tasks.

However, there are other aspects of our systems that we consider relevant. First of all, they are completely independent of the domain because they only relies on the terms present in the training set. That is to say, they do not require a costly process of feature engineering or very complex hand-crafted features specific of the problem under consideration. That independence was evident in this Lab where only a parameter setting was carried out on one of the data sets (depression) and the same configuration was used in the other one (anorexia). The excellent results obtained in both cases provide strong evidence of this independence and robustness. Another aspect that deserves special attention is that both approaches use very simple rules to decide when to stop reading and classify a user as positive. That contrasts with other approaches that require very complex and difficult to understand methods to make those decisions.

As future work we plan to extend the use of FTVT and SIC to other ERD problems such as the identification of sexual predators, people with suicide tendency and early rumour detection. In those cases, we consider that the ease and simplicity that our methods provide to be migrated from one domain to another make these applications a rather trivial process.

References

1. Marcelo L. Errecalde, Ma. Paula Villegas, Dario G. Funez, Ma. José Garciarena Ucelay, and Leticia C. Cagnina. Temporal variation of terms as concept space for early risk prediction. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Vol 1866*, 2017.

2. Zhixing Li, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441–448, February 2011.
3. Adrián Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Efstathios Stamatatos. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 89:134 – 147, 2015.
4. David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.
5. David E Losada, Fabio Crestani, and Javier Parapar. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer, 2017.
6. David E. Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk – Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, Avignon, France, 2018.
7. Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1):40–48, 2010.