# Temporal Variation of Terms as concept space for early risk prediction

Marcelo L. Errecalde[1], Ma. Paula Villegas[1], Dario G. Funez[1], Ma. José Garciarena Ucelay[1], and Leticia C. Cagnina[1,2]

[1] LIDIC Research Group, Universidad Nacional de San Luis, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
{merrecalde,villegasmariapaula74,funezdario}@gmail.com
{mjgarciarenaucelay,lcagnina}@gmail.com

**Abstract.** Early risk prediction involves three different aspects to be considered when an automatic classifier is implemented for this task: a) support for classification with partial information read up to different time steps, b) support for dealing with unbalanced data sets and c) a policy to decide *when* a document could be classified as belonging to the relevant class with a reasonable confidence. In this paper we propose an approach that naturally copes with the first two aspects and shows good perspectives to deal with the last one. Our proposal, named *temporal variation of terms* (TVT) is based on using the variation of vocabulary along the different time steps as concept space to represent the documents. Results with the eRisk 2017 data set show a better performance of TVT in comparison to other successful semantic analysis approaches and the standard BOW representation. Besides, it also reaches the best reported results up to the moment for $ERDE_5$ and $ERDE_{50}$ error evaluation measures.
**Keywords:**Early Risk Detection, Unbalanced Data Sets, Text Representations, Semantic Analysis Techniques.

## 1 Introduction

Early risk detection (ERD) is a new research area potentially applicable to a wide variety of situations such as detection of potential paedophiles, people with suicidal inclinations, or people susceptible to depression, among others. In a ERD scenario, data are sequentially read as a stream and the challenge consists in detecting risk cases as soon as possible. A usual situation in these cases is that the target class (the risky one) is clearly under-sampled with respect to the control class (the non-risky one). That unequal distribution between the positive (minority) class and the negative one, is a well-known problem in categorization tasks and popularly referred as *unbalanced data sets* (UDS).

Besides dealing with the UDS problem, an ERD system needs to consider the problem of assigning a class to documents when only partial information is available. A document is processed as a sequence of terms, and the goal is to devise a method that can make predictions with the information read up to a

specific point of the time. That aspect, that could be named as *classification with partial information* (CPI) might be addressed with a simple approach that consists in training with complete documents as usual and considering the partial documents read up to the classification point as standard "complete" documents. In [3] the CPI aspect was considered by analysing the robustness of the Naïve Bayes algorithm to deal with partial information.

Last, but not least, an ERD system needs to consider not only which class should be assigned to a document, but also deciding *when* to make that assignment. This aspect, that we will refer as the *classification time decision* (CTD) issue has been addressed with very simple heuristic rules[3] although more elaborated approaches might be used.

In this article we propose an original idea that explicitly considers the sequentiality of data to deal with the unbalanced data sets problem. In a nutshell, we use the temporal variation of terms as concept space of a recent concise semantic analysis (CSA) approach [7]. CSA is an interesting document representation technique which models words and documents in a small "concept space" whose concepts are obtained from category labels. CSA has obtained good results in author profiling tasks [8] and the variant proposed in this article, named *temporal variation of terms* (TVT), seems to show some interesting characteristics to deal with the ERD problem. In fact, it obtained a robust performance on the eRisk 2017 data set and reached the best (lowest) reported results up to the moment for $ERDE_5$ and $ERDE_{50}$ error evaluation measures.

The rest of this document is organized as follows: Section 2 describes our proposed method for the ERD problem. Section 3 shows the obtained results with our method on the eRisk 2017 dataset. Finally, Section 4 depicts potential future works and the obtained conclusions.

## 2 The proposed method

Our method is based on the concise semantic analysis (CSA) technique proposed in [7] and later extended in [8] for author profiling tasks. Therefore, we first present in Subsection 2.1 the key aspects of CSA and then explain in Subsection 2.2 how we instantiate CSA with concepts derived from the terms used in the temporal chunks analysed by an ERD system at different time steps.

### 2.1 Concise Semantic Analysis

Standard text representation methods such as Bag of Words (BoW) suffer of two well known drawbacks. First, their high dimensionality and sparsity; second, they do not capture relationships among words. CSA is a semantic analysis technique that aims at dealing with those shortcomings by interpreting words and documents in a space of *concepts*. Differently from other semantic analysis approaches such as *latent semantic analysis* (LSA) [2] and *explicit semantic*

[3] For instance, exceeding a specific confidence threshold in the prediction of the classifier [9].

*analysis* (ESA) [4] which usually require huge computing costs, CSA 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).

To explain the main concepts of the CSA technique we first introduce some basic notation that will be used in the rest of this work. Let $\mathcal{D} = \{\langle d_1, y_1\rangle, \ldots, \langle d_n, y_n\rangle\}$ be a training set formed by $n$ pairs of documents ($d_i$) and variables ($y_i$) that indicate the concept the document is associated with, $y_i \in \mathcal{C}$ where $\mathcal{C} = \{c_1, \ldots, c_q\}$ is the *concept space*. For the moment, consider that these concepts correspond to standard category labels although, as we will see later, they might represent more elaborate aspects. In this context, we will denote as $\mathcal{V} = \{t_1, \ldots, t_m\}$ to the vocabulary of terms of the collection being analysed.

**Representing terms in the concept space** In CSA, each term $t_i \in \mathcal{V}$ is represented as a vector $\mathbf{t_i} \in \mathbb{R}^q$, $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q}\rangle$. Here, $t_{i,j}$ represents the degree of association between the term $t_i$ and the concept $c_j$ and its computation requires some basic steps that are explained below. First, the raw term-concept association between the $i$th term and the $j$th concept, denoted $w_{ij}$, will be obtained. If $D_{c_u} \subseteq \mathcal{D}$, $D_{c_u} = \{d_r \mid \langle d_r, y_s\rangle \in \mathcal{D} \wedge y_s = c_u\}$ is the subset of the training instances whose label is the concept $c_u$, then $w_{ij}$ might be defined as:

$$w_{ij} = \sum_{\forall d_k \in D_{c_j}} \log_2\left(1 + \frac{tf_{ik}}{len(d_k)}\right) \tag{1}$$

where $tf_{ik}$ is the number of occurrences of the term $t_i$ in the document $d_k$ and $len(d_k)$ is the length (number of terms) of $d_k$.

As noted in [7] and [8], direct use of $w_{ij}$ to represent terms in the vector $\mathbf{t_i}$ could be sensible to highly unbalanced data. Thus, some kind of normalization is usually required and, in our case, we selected the one proposed in [8]:

$$t'_{ij} = \frac{w_{ij}}{\sum_{i=1}^{m} w_{ij}} \tag{2}$$

$$t_{ij} = \frac{t'_{ij}}{\sum_{j=1}^{q} w_{ij}} \tag{3}$$

With this last conversion we finally obtain, for each term $t_i \in \mathcal{V}$, a $q$-dimensional vector $\mathbf{t_i}$, $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q}\rangle$ defined over a space of $q$ concepts. Up to now, those concepts correspond to the original categories used to label the documents. Later, we will use other more elaborated concepts.

**Representing documents in the concept space** Once the terms are represented in the $q$-dimensional concept space, those vectors can be used to represent documents in the same concept space. In CSA, documents are represented as the

central vector of all the term vectors they contain [7]. Terms have different importance for different documents so it is not a good idea computing that vector for the document as the simple average of all its term vectors. Previous works in BoW [6] have considered different statistic techniques to weight the importance of terms in a document such as $tfidf$, $tfig$, $tf\chi^2$ or $tfrf$, among others. Here, we will use the approach used in [8] for author profiling that represents each document $d_k$ as the weighted aggregation of the representations (vectors) of terms that it contains:

$$\mathbf{d_k} = \sum_{t_i \in d_k} \left( \frac{tf_{ik}}{len(d_k)} \times \mathbf{t_i} \right) \tag{4}$$

Thus, documents are also represented in a $q$-dimensional concept space (i.e., $\mathbf{d_k} \in \mathbb{R}^q$) which is much smaller in dimensionality than the one required by standard BoW approaches ($q \ll m$).

### 2.2 Temporal Variation of Terms

In Subsection 2.1 we said that the concept space $\mathcal{C}$ usually corresponds to standard category names used to label the training instances in supervised text categorization tasks. In this scenario, that in [7] is referred as *direct derivation*, each category label simply corresponds to a concept. However, in [7] also are proposed other alternatives like *split derivation* and *combined derivation*. The former uses the low-level labels in hierarchical corpora and the latter is based on combining semantically related labels in a unique concept. In [8] those ideas are extended by first clustering each category of the corpora and then using those subgroups (sub-clusters) as new concept space.[4]

As we can see, the common idea to all the above approaches is that once a set of documents is identified as belonging to a group/category, that category can be considered as a concept and CSA can be applied in the usual way. We take a similar view to those works by considering that the positive (minority) class in ERD problems can be augmented with the concepts derived from the sets of partial documents read along the different time steps. In order to understand this idea it is necessary to first introduce a sequential work scheme as the one proposed in [9] for research in ERD systems for depression cases.

Following [9], we will assume a corpus of documents written by $p$ different individuals ($\{I_1, \ldots, I_p\}$). For each individual $I_l$ ($l \in \{1, \ldots, p\}$), the $n_l$ documents that he has written are provided in chronological order (from the oldest text to the most recent text): $D_{I_l,1}, D_{I_l,2}, \ldots, D_{I_l,n_l}$. In this context, given these $p$ streams of messages, the ERD system has to process every sequence of messages (in the chronological order they are produced) and to make a binary decision (as early as possible) on whether or not the individual might be a positive case of depression. Evaluation metrics on this task must be time-aware, so an early risk detection error (ERDE) is proposed. This metric not only takes into account the

[4] In that work, concepts are referred as *profiles* and subgroups as *sub-profiles*.

correctness of the (binary) decision but also the delay taken by the system to make the decision.

In a usual supervised text categorization task, we would only have two category labels: *positive* (risk/depressive case) and *negative* (non-risk/non-depressive case). That would only give two concepts for a CSA representation. However, in ERD problems there is additional temporal information that could be used to obtain an improved concept space. For instance, the training set could be split in $h$ "chunks", $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_h$, in such a way that $\hat{C}_1$ contains the oldest writings of all users (first $(100/h)\%$ of submitted posts or comments), chunk $\hat{C}_2$ contains the second oldest writings, and so forth. Each chunk $\hat{C}_k$ can be partitioned in two subsets $\hat{C}_k^+$ and $\hat{C}_k^-$, $\hat{C}_k = \hat{C}_k^+ \bigcup \hat{C}_k^-$ where $\hat{C}_k^+$ contains the positive cases of chunk $\hat{C}_k$ and $\hat{C}_k^-$ the negatives ones of this chunk.

It is interesting to note that we can also consider the data sets that result of concatenating chunks that are contiguous in time and using the notation $\hat{C}_{i-j}$ to refer to the chunk obtained from concatenating all the (original) chunks from the $i$th chunk to the $j$th chunk (inclusive). Thus, $\hat{C}_{1-h}$ will represent the data set with the complete streams of messages of all the $p$ individuals. In this case, $\hat{C}_{1-h}^+$ and $\hat{C}_{1-h}^-$ will have the obvious semantic specified above for the complete documents of the training set.

The classic way of constructing a classifier would be to take the complete documents of the $p$ individuals ($\hat{C}_{1-h}$) and use an inductive learning algorithm such as SVM or Naïve Bayes to obtain that classifier. As we mentioned earlier, another important aspect in EDS systems is that the classification problem being addressed is usually highly unbalanced (UDS problem). That is, the number of documents of the majority/negative class ("non-depression") is significantly larger than that of the minority/positive class ("depression"). More formally, following the previously specified notation $\mid \hat{C}_{1-h}^- \mid \gg \mid \hat{C}_{1-h}^+ \mid$.

An alternative to try to alleviate the UDS problem would be to consider that the minority class is formed not only by the complete documents of the individuals but also by the partial documents obtained in the different chunks. Following the general ideas posed in CSA, we could consider that the partial documents read in the different chunks represent "temporal" concepts that should be taken into account. In this context, one might think that variations of the terms used in these different sequential stages of the documents may have relevant information for the classification task. With this idea in mind, the method proposed in this work named *temporal variation of terms* (TVT) arises, which consists in enriching the documents of the minority class with the partial documents read in the first chunks. These first chunks of the minority class, along with their complete documents, will be considered as a new concept space for a CSA method.

Therefore, in TVT we first determine the number $f$ of initial chunks that will be used to enrich the minority (positive) class. Then, we use the document sets $\hat{C}_1^+, \hat{C}_{1-2}^+, \ldots, \hat{C}_{1-f}^+$ and $\hat{C}_{1-h}^+$ as concepts for the positive class and $\hat{C}_{1-h}^-$ for the negative class. Finally, we represent terms as documents in this new $(f+2)$-dimensional space using the CSA approach explained in Section 2.1.

## 3 Experimental Analysis

### 3.1 Data Set

Our approach was tested on the data set used in the eRisk 2017 pilot task[5] and described in [9]. 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, the 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. This collection was split into a training and a test set that we will refer as $\mathcal{TR}_{DS}$ and $\mathcal{TE}_{DS}$ respectively. The (training) $\mathcal{TR}_{DS}$ set contained 486 users (83 positive, 403 negative) and the (test) $\mathcal{TE}_{DS}$ set contained 401 users (52 positive, 349 negative). The users labeled as positive are those that have explicitly mentioned that they have been diagnosed with depression.

This task was divided into a training stage and a testing stage. In the first one, the participating teams had access to the $\mathcal{TR}_{DS}$ set with all chunks of all training users. They could therefore tune their systems with the training data. To reproduce the same conditions of the pilot task, we use the training set ($\mathcal{TR}_{DS}$) to generate a new corpus divided into a training set (that we will refer as $\mathcal{TR}_{DS} - train$) and a test set (named $\mathcal{TR}_{DS} - test$) with the same categories (depressed and non-depressed) for each sequence of writings of the users in the collection. Those sets maintained the same proportions of post per user and words per user as described in [9]. $\mathcal{TR}_{DS} - train$ and $\mathcal{TR}_{DS} - test$ were generated by randomly selecting around a 70% of writings for the first one and the rest 30% for the second one. Thus, $\mathcal{TR}_{DS} - train$ resulted in 351 writings (63 positive, 288 negative) meanwhile $\mathcal{TR}_{DS} - test$ contains 135 individuals (20 positive, 115 negative). In the pilot task the collection of writings was divided into 10 chunks, so we made the same division on $\mathcal{TR}_{DS} - train$ and $\mathcal{TR}_{DS} - test$.

### 3.2 Experimental Results

We reproduced the same conditions faced by the participants of the eRisk pilot task, so we first worked on the data set released on the training stage ($\mathcal{TR}_{DS}$) and then, the obtained models were tested on the test stage ($\mathcal{TE}_{DS}$). The activities carried out at each stage are described below.

**Training stage** CSA is a document representation that aims at addressing some drawbacks of classical representations such as BoW. On the other hand, TVT is supposed to extend CSA by defining concepts that capture the sequential aspects of the ERD problems and the variations of vocabulary observed in the distinct stages of the individuals' writings. Thus, CSA and BoW arise as obvious candidates to compare TVT in the data set used in the pilot task. Those three

[5] `http://early.irlab.org/task.html`

representations were evaluated with different learning algorithms such as SVM, Naïve Bayes and Random Forest, among others. In each case, the best parameters were selected for each algorithm-representation combination (*model*) and the reported results correspond to the best obtained values.

We tested BoW with different weighting schemes and learning algorithms but, in all cases, the best results were obtained with binary representations and the Naïve Bayes algorithm. From now on, all references to "BoW" will stand for that setting. We use CSA with representations of terms with normalized weights according to Equations 2 and 3 and document representations obtained from Equation 4 as proposed in [8] for author profiling tasks. We named this setting as $CSA^{\star}$. For the TVT representation, a decision must be made related to the number $f$ of chunks that will enrich the minority (positive) class. In our studies, we use $f = 4$ and, in consequence, the positive class was represented by 5 concepts. In that way, the number of documents in the "depressed" class was incremented by 5 with respect to the original size, from 83 positive instances to 415. As we can see, with this technique we are also obtaining some kind of "balancing" in the size of both classes and addressing in that way another usual problem that we previously refer as the UDS problem.

A particularity that ERD methods must consider is the criterion used to decide *when* (in what situations) the classification generated by the system is considered the final/definitive decision on the evaluated instances (the *classification time decision* (CTD) issue). We will start our evaluation of the different document representations and algorithms assuming that the classification is made on a static "chunk by chunk" basis. That is, for each chunk $\hat{C}_i$ provided to the ERD systems we will evaluate their performance considering that all the models are (simultaneously) applied to the writings received up to the chunk $\hat{C}_i$. With this kind of information it will be possible to observe to what extent the different approaches are robust to the partial information in the different stages, in which moment they start to obtain acceptable results, and other interesting statistics.

Tables 1, 2 and 3 show the results of experiments for this static "chunk by chunk" classification scheme. Values of *precision* ($\pi$), *recall* ($\rho$) and $F_1$-measure ($F_1$) of the target ("depressed") class are reported for each considered model. Statistics also include the *early risk detection error* (ERDE) measure proposed in [9]. This measure considers 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 [9].

In our study we consider the two values of $o$ used in the pilot task: $o = 5$ ($ERDE_5$) and $o = 50$ ($ERDE_{50}$). In each chunk, classifiers usually produce their predictions with some kind of "confidence", in general, the estimated probability of the predicted class. In those cases, we can select different thresholds $tr$ considering that an instance (document) is assigned to the target class when its associated probability $p$ is greater (or equal) than certain threshold $tr$ ($p \geq tr$). In this study we evaluated 5 different settings for the probabilities assigned for each classifier: $p = 1$, $p \geq 0.9$, $p \geq 0.8$, $p \geq 0.7$ and $p \geq 0.6$. Due to space constraints, only the best results obtained with a particular setting are shown.[6]

Table 1 shows the results obtained with a BoW representation and a Naïve Bayes classifier. Those values correspond to the setting where an instance is considered as depressive if the classifier assigns to the target/positive class a probability greater or equal than 0.8 ($p \geq 0.8$). Surprisingly, the best results for all the considered measures are obtained on the first chunk. In this chunk, we can observe that this model only recovers a 45% of the depressed individuals. However, this is not the worst aspect. Only a 12% of the individual classified as "depressed" effectively had this condition resulting in consequence in a very low $F_1$ measure (0.19). Table 2 shows similar results when a $CSA^\star$-RF (random forest) combination with $p \geq 0.6$ is used to classify the writings of the individuals. Here, $F_1$ measure is also low but we can observe a deterioration in the ($ERDE_5$) and ($ERDE_{50}$) error values with respect to the previous model.

Finally, in Table 3, the results of TVT with a Naïve Bayes algorithm and $p \geq 0.6$ are shown. There, we can see a remarkable improvement in the performance of the classifier in the chunk 3 with excellent values of $ERDE_{50}$ (7.02), precision $\pi$ (0.63), recall $\rho$ (0.85) and $F_1$ measure (0.72). Analysing the results along the 10 considered chunks we observe how the measures keep improving from the chunk 1 up to reach the best values in chunk 3 and, from then on, they start to deteriorate chunk by chunk and obtaining the worst results on the last two chunks. As weak points of those results we can say that the best value of $ERDE_5$ obtained in chunk 1 is not very good. Besides, even thought $ERDE_{50}$ values are acceptable for most of the considered chunks, they need at least two chunks to show a competitive performance. That aspect looks reasonable if we consider that TVT is based on the variation of terms between consecutive chunks and that information is not available on the first chunk.

As general conclusion to the "chunk by chunk" analysis, we could say that imbalanced classes seem to affect in a different way to the different methods. BoW and CSA directly depend on the vocabulary of positive and negative classes. In the first chunk where texts are supposed to be the shortest, relevant words of the positive class appearing in the posts will probably have more chance of being "balanced" with respect to the words appearing in the negative class. That makes classifiers be more sensitive to the positive class and, in consequence, the recall and general performance is improved. As more information is read, words related to the negative class are more probable to occur introducing "noise" and

[6] All the tables generated for the different probabilities can be downloaded from `https://sites.google.com/site/lcagnina/research/Tables_eRisk17.rar`

**Table 1.** *Model*: *BoW* + Naïve Bayes ($p \geq 0.8$). "Chunk by chunk" setting. $ERDE_5$, $ERDE_{50}$, $F_1$-measure ($F_1$), precision ($\pi$) and recall ($\rho$) of the "depressed class".

| | $ch_1$ | $ch_2$ | $ch_3$ | $ch_4$ | $ch_5$ | $ch_6$ | $ch_7$ | $ch_8$ | $ch_9$ | $ch_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $ERDE_5$ | **18.09** | 20.98 | 21.5 | 21.73 | 21.95 | 21.95 | 21.95 | 21.95 | 22.17 | 22.17 |
| $ERDE_{50}$ | **15.17** | 16.84 | 20.77 | 20.25 | 21.21 | 21.95 | 21.95 | 21.52 | 22.17 | 22.17 |
| $F_1$ | **0.19** | 0.16 | 0.09 | 0.11 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.13 |
| $\pi$ | **0.12** | 0.11 | 0.06 | 0.17 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.08 |
| $\rho$ | **0.45** | 0.35 | 0.2 | 0.25 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 |

**Table 2.** *Model*: $CSA^\star$ + RF ($p \geq 0.6$). "Chunk by chunk" setting. $ERDE_5$, $ERDE_{50}$, $F_1$-measure ($F_1$), precision ($\pi$) and recall ($\rho$) of the "depressed class".

| | $ch_1$ | $ch_2$ | $ch_3$ | $ch_4$ | $ch_5$ | $ch_6$ | $ch_7$ | $ch_8$ | $ch_9$ | $ch_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $ERDE_5$ | **21.93** | 25.64 | 25.46 | 25.57 | 26.12 | 25.68 | 25.68 | 25.46 | 25.35 | 25.68 |
| $ERDE_{50}$ | **19.47** | 24.94 | 25.46 | 23.35 | 25.37 | 24.2 | 23.46 | 22.5 | 22.39 | 23.47 |
| $F_1$ | **0.19** | 0.08 | 0.05 | 0.1 | 0.06 | 0.08 | 0.13 | 0.16 | 0.16 | 0.14 |
| $\pi$ | **0.11** | 0.05 | 0.03 | 0.06 | 0.04 | 0.05 | 0.07 | 0.09 | 0.09 | 0.08 |
| $\rho$ | **0.6** | 0.25 | 0.15 | 0.3 | 0.2 | 0.25 | 0.4 | 0.5 | 0.5 | 0.45 |

affecting in consequence the performance. TVT does not seem to be so affected by this problem showing a more stable performance along all the chunks, with the best results in the third chunk and then with a little deterioration from then on. Those results could be giving evidence that the variation of terms (with $f = 4$) allows to better detect the occurrence of relevant words of the positive class in the first chunks. However, it also seems to be affected by the unbalance problem in subsequent chunks, although in a lower level than BoW and CSA representations. Unfortunately, verifying those hypotheses would require considering "more balanced" settings and different $f$ values what is out of the scope of this paper. However, that important aspect will be addressed in future works

Another approach for the CTD issue could be directly use the probability (or some measure of confidence) assigned by the classifier to decide *when* to stop reading a document and giving its classification. That approach, that in [9] is referred as *dynamic*, only considers that this probability exceeds some particular threshold to classify the instance/individual as positive. That means, that different streams of messages could be classified as "depressed" in different stages (chunks). Table 4 show those statistics for BoW, $CSA^\star$ and TVT representations for those learning algorithms and probability thresholds that obtained the best performance. There, we can see that TVT representation, with a Naïve Bayes and classifying instances as depressed when the assigned probability is 1, obtains the best results for the measures we are more interested in: $ERDE_5$, $ERDE_{50}$ and $F_1$-measure. In this context, BoW gets a better recall value but at the expense of lowering the precision values resulting in a poor $F_1$-measure.

**Testing stage** The previous results were obtained by training the classifiers with the $\mathcal{TR}_{DS} - train$ data set and testing them with the $\mathcal{TR}_{DS} - test$ data

**Table 3.** *Model*: $TVT$ + Naïve Bayes ($p \geq 0.6$). "Chunk by chunk" setting. $ERDE_5$, $ERDE_{50}$, $F_1$-measure ($F_1$), precision ($\pi$) and recall ($\rho$) of the "depressed class".

| | $ch_1$ | $ch_2$ | $ch_3$ | $ch_4$ | $ch_5$ | $ch_6$ | $ch_7$ | $ch_8$ | $ch_9$ | $ch_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $ERDE_5$ | **14.24** | 14.27 | 14.59 | 14.83 | 15.17 | 15.51 | 15.74 | 15.84 | 16.21 | 16.13 |
| $ERDE_{50}$ | 10.80 | 7.22 | **7.02** | 9.24 | 9.25 | 9.97 | 10.73 | 10.73 | 11.06 | 10.96 |
| $F_1$ | 0.42 | 0.65 | **0.72** | 0.67 | 0.67 | 0.67 | 0.64 | 0.64 | 0.57 | 0.58 |
| $\pi$ | 0.39 | 0.58 | **0.63** | 0.60 | 0.60 | 0.60 | 0.58 | 0.58 | 0.50 | 0.52 |
| $\rho$ | 0.45 | 0.75 | **0.85** | 0.75 | 0.75 | 0.75 | 0.70 | 0.70 | 0.65 | 0.65 |

**Table 4.** *Dynamic Models* for BoW-NB, $CSA^\star$-NB and TVT-NB.

| | $ERDE_5$ | $ERDE_{50}$ | $F_1$ | $\pi$ | $\rho$ |
|---|---|---|---|---|---|
| BoW ($p \geq 0.8$) | 21.05 | 18.13 | 0.24 | 0.14 | **0.75** |
| $CSA^\star$-NB($p = 1$) | 23.09 | 23.07 | 0.06 | 0.04 | 0.15 |
| TVT-NB ($p = 1$) | **14.13** | **11.25** | **0.40** | **0.47** | 0.35 |

set. The obvious question now is if similar results are obtained by training with the full training set of the pilot task ($\mathcal{TR}_{DS}$) and using the classifiers with the data set $\mathcal{TE}_{DS}$ that was incrementally released during the testing phase of the pilot task. In this new scenario, the TVT representation was used with a simple rule for the CTD issue that consists in classifying all the individual in the chunk 3 as positive (depressed) if a Naïve Bayes classifier produced a probability equal or greater than 0.6 for the positive class. That strategy, that we will refer as $TVT^3_{p\geq 0.6}$, is motivated by the good results showed by TVT in Table 3. We also tested the BoW, $CSA^\star$ and TVT representations with dynamic strategies and using those probabilities that best values obtained in the training stage. As baselines we also tested two approaches described in [9] that will be named as *Ran* and $Min$. *Ran*, simply emits a random decision ("depressed"/"non-depressed") for each user in the first chunk. $Min$, on the other hand, stands for "minority" and consists in classifying each user as "depressive" in the first chunk.

Table 5 shows the performance of all the above mentioned approaches on the test set of the pilot task ($\mathcal{TE}_{DS}$). We also included the results reported in the eRisk page for the systems that obtained the best $ERDE_5$ ($FHDO-BCSGB$), $ERDE_{50}$ ($UNSLA$) and $F_1$ ($FHDO-BCSGB$) measures on the pilot task. Here we can observe that results obtained with $TVT^3_{p\geq 0.6}$ are not as good as those obtained in the training stage. However, the setting TVT-NB ($p = 1$) would have obtained the best $ERDE_5$ score and the third $ERDE_{50}$ value, with a small difference respect to the best reported value (9.84 versus 9.68).

Those good results of TVT were achieved taking into account the *best* parameters obtained in the training stage. However, it also would be interesting analysing what would have been the TVT's performance if other parameter settings had been selected. Table 6 shows this type of information by reporting the results obtained with different learning algorithms (Naïve Bayes and Random Forest) and different probability values for "dynamic" approaches to the CTD aspect. The results are conclusive in this case. TVT shows a high robustness in

**Table 5.** Results on the $\mathcal{TE}_{DS}$ test set.

| | $ERDE_5$ | $ERDE_{50}$ | $F_1$ | $\pi$ | $\rho$ |
|---|---|---|---|---|---|
| *Ran* | 16.83 | 14.63 | 0.17 | 0.11 | 0.4 |
| *Min* | 21.67 | 15.03 | 0.23 | 0.13 | 1 |
| BoW ($p \geq 0.8$) | 16.45 | 10.87 | 0.38 | 0.25 | 0.77 |
| $CSA^{\star}$-NB($p = 1$) | 20.58 | 19.58 | 0.05 | 0.03 | 0.15 |
| $TVT^{3}_{p \geq 0.6}$ | 13.64 | 10.17 | 0.53 | 0.46 | 0.62 |
| TVT-NB ($p = 1$) | **12.38** | 9.84 | 0.42 | 0.50 | 0.37 |
| $FHDO - BCSGA$ | 12.82 | 9.69 | **0.64** | 0.61 | 0.67 |
| $FHDO - BCSGB$ | 12.70 | 10.39 | 0.55 | **0.69** | 0.46 |
| *UNSLA* | 13.66 | **9.68** | 0.59 | 0.48 | 0.79 |

**Table 6.** Results of TVT with different learning algorithms and probability values.

| | $ERDE_5$ | $ERDE_{50}$ | $F_1$ | $\pi$ | $\rho$ |
|---|---|---|---|---|---|
| TVT-NB ($p \geq 0.6$) | 13.59 | 8.40 | 0.50 | 0.37 | **0.75** |
| TVT-NB ($p \geq 0.7$) | 13.43 | 8.24 | 0.51 | 0.39 | **0.75** |
| TVT-NB ($p \geq 0.8$) | 13.13 | **8.17** | 0.54 | 0.42 | 0.73 |
| TVT-NB ($p \geq 0.9$) | 13.07 | 8.35 | 0.52 | 0.42 | 0.69 |
| TVT-NB($p = 1$) | 12.38 | 9.84 | 0.42 | 0.50 | 0.37 |
| TVT-RF ($p \geq 0.6$) | 12.46 | 8.37 | 0.55 | 0.49 | 0.63 |
| TVT-RF ($p \geq 0.7$) | 12.49 | 8.52 | 0.55 | 0.50 | 0.62 |
| TVT-RF ($p \geq 0.8$) | **12.30** | 8.95 | **0.56** | 0.54 | 0.58 |
| TVT-RF ($p \geq 0.9$) | 12.34 | 10.28 | 0.47 | 0.55 | 0.40 |
| TVT-RF($p = 1$) | 12.82 | 11.82 | 0.20 | **0.67** | 0.12 |

the $ERDE$ measures *independently* of the algorithm used to learn the model and the probability used in the dynamic approaches. Most of the $ERDE_5$ values are low and in 7 out of 10 settings the $ERDE_{50}$ values are lowest than the best reported in the pilot task (*UNSLA*: 9.68). In this context, TVT achieves the best reported $ERDE_5$ value up to now (12.30) with the setting TVT-RF ($p \geq 0.8$) and the lowest $ERDE_{50}$ value (8.17) with the model TVT-NB ($p \geq 0.8$).

## 4 Conclusions and future work

In this article we present *temporal variation of terms* (TVT) an approach for early risk detection based on using the variation of vocabulary along the different time steps as concept space for document representation. TVT naturally copes with the sequential nature of ERD problems and also gives a tool for dealing with unbalanced data sets. Preliminary results with the eRisk 2017 data set show a better performance of TVT in comparison to other successful semantic analysis approach and the standard BOW representation. It also shows a robust performance along different parameter settings and reaches the best reported results up to the moment for $ERDE_5$ and $ERDE_{50}$ error evaluation measures.

As future work, we plan to apply the TVT approach to other problems that can be directly tackled as ERD problems such as sexual predation and suicide

discourse identification. Our first option to work will be the corpus used in the PAN-2012 competition on sexual predator identification [5] which shares several characteristics with the data set used in the present work such as the sequentially of data, unbalanced classes and the requirement of detecting the minority class (predator) as soon as possible, among others.

TVT is explicitly based on the enrichment of the minority class with new concepts derived from the partial information obtained from the initial chunks. However, some improvements can be achieved by also clustering the negative class as proposed by [8] in author profiling tasks. We carried out some initial experiments by combining TVT with the clustering of the negative class but more study is required to determine how both approaches can be effectively integrated. Besides, in the present work, the election of $f = 4$ mainly aimed at obtaining balanced positive and negative classes. In future works, different $f$ values will be considered to see how they impact on the TVT's performance.

TVT provides, as a side effect, an interesting tool for dealing with the unbalanced data set problem. We plan to apply TVT on unbalanced data sets that do not necessarily correspond to the ERD field and comparing it against other well known methods in this area, such as SMOTE [1]. Finally, it would be interesting comparing the concept space used in our approach against other recent and effective representations based on word embeddings. In this context, it could also be analysed how our concept space representation can be extended/improved with information provided by those embeddings.

## References


1. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. Ph. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.
2. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the ASIS*, 41(6):391–407, 1990.
3. H. Jair Escalante, M. Montes-y-Gómez, L. Villaseñor Pineda, and M. Errecalde. Early text classification: a naïve solution. In A. Balahur, E. Van der Goot, P. Vossen, and A. Montoyo, editors, *Proc. of WASSA@NAACL-HLT 2016, 2016, San Diego, California, USA*, pages 91–99. The Association for Computer Linguistics, 2016.
4. E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. *JAIR*, 34(1):443–498, March 2009.
5. G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, pages 1–12, 2012.
6. M. Lan, Ch. Tan, J. Su, and Y. Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE TPAMI*, 31(4):721–735, 2009.
7. Z. Li, Z. Xiong, Y. Zhang, Ch. Liu, and K. Li. Fast text categorization using concise semantic analysis. *Pattern Recogn. Lett.*, 32(3):441–448, February 2011.
8. A. Pastor López-Monroy, M. Montes y Gómez, H. Jair Escalante, L. Villaseñor-Pineda, and E. Stamatatos. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 89:134 – 147, 2015.
9. D. E. Losada and F. Crestani. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th Int. Conf. of the CLEF Association, Portugal*, pages 28–39, 2016.