LIDIC - UNSL's participation at eRisk 2017: Pilot task on Early Detection of Depression Notebook for eRisk at CLEF 2017

Ma. Paula Villegas¹, Dario G. Funez¹, Ma. José Garciarena Ucelay¹, Leticia C. Cagnina^{1,2}, and Marcelo L. Errecalde¹

 ¹ LIDIC Research Group, Universidad Nacional de San Luis, Argentina
² Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) {villegasmariapaula74,funezdario,mjgarciarenaucelay}@gmail.com {lcagnina,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 eRisk 2017 pilot task. The main goal of this task is considering early risk detection scenarios, depression in this case, where the issue of getting timely predictions with a reasonable confidence level becomes critical. In the pilot task, systems must be able to sequentially process pieces of evidence and detect early traces of depression as soon as possible. The used data set is a collection of writings labeled as "depressed" or "non-depressed" that were released to the pilot task participants in two different stages: training and test. Our proposal for this task was based on 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. That "temporal" approach was complemented with other standard text categorization models in some specific situations that seemed not to be correctly addressed by our approach alone. In the test stage, the resulting system obtained the lowest $ERDE_{50}$ error on a total of 30 submissions from 8 different institutions. However, once the "golden-truth" of the testing data set was released, we could verify that our "temporal" approach alone might have obtained very robust results and the lowest reported ERDE error for both thresholds used in the pilot task.

Keywords: Early Risk Detection, Unbalanced Data Sets, Text Representations, Semantic Analysis Techniques.

1 Introduction

The widespread use of Internet, social networks and other computer technologies has marked the beginning of a new era in the communication among people. In this context, we have now available a lot of information that might be useful to detect, in a timely way, those situations that might be potentially dangerous or risky for the physical well-being and health of individuals, communities and social organizations. As examples of those situations we can mention the detection of potential paedophiles, people with suicidal inclinations, or people susceptible to depression, among others.

This type of situations have started to be studied in a new research field known as *early risk detection* (ERD) which has received increasingly interest from scientific researchers at world level due to the important impact that it could have in many relevant and current problems of the real world. In this context, this year was organized the first early risk prediction conference eRisk 2017^3 in the context of the CLEF 2017 Workshop. As part of this event, a pilot task was organized and the present article describes our participation in this task.

The eRisk 2017 pilot task was organized into two different stages: training and test stages. It also assumed an ERD scenario, that is, data are sequentially read as a stream and the challenge consists in detecting risk cases as soon as possible. In order to reproduce this scenario, the eRisk 2017's organizers released the test data set following a sequential, "chunk by chunk" criterion; that is, the first chunk contained the oldest 10% of the messages, the second chunk contains the second oldest 10%, and so forth up to complete 10 chunks that represent the full writing of the analysed individuals.

To deal with this problem we implemented an original proposal that we named temporal variation of terms (TVT) [3]. However, at the training stage, TVT seemed to show some weakness at specific chunks so we decided to complement it with other standard methods to help our approach in these specific situations. Thus, our implemented ERD system is in fact a combined system that strongly depends on TVT but also uses other methods as additional source of "opinions". The resulting system had a very acceptable performance on the data set released in the test stage obtaining the lowest $ERDE_{50}$ error on a total of 30 submissions from 8 different institutions. However, once the "golden-truth" of the test data set was released, we could verify that TVT alone might have obtained very robust results and the lowest reported ERDE error for both thresholds used in the pilot task. For this reason, we also include an additional section in this work with preliminary results obtained with TVT alone on the test set in order to observe the potential of this method for general ERD problems.

The rest of the article is organized as follows: Section 2 describes some general aspects of the data set used in the pilot task and the methods used in our ERD system. Next, in Section 3 the activities carried out in the training stage are described and the justifications of the main design decisions made on our ERD system, are presented. Section 4 shows the obtained results with our proposal on the eRisk 2017 dataset released in the test stage. Then, some complementary results are presented in Section 5 where interesting aspects are shown on the performance of TVT working alone on the test set. Finally, Section 6 depicts potential future works and the obtained conclusions.

³ http://early.irlab.org/

2 Data set and methods

2.1 Data Set

The data set used in the pilot task of the eRisk competition⁴ was initially described in [8]. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, "depressed" and "nondepressed" 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. 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 \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.

The pilot 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 \mathcal{TR}_{DS} data set with all chunks of all training users. They could therefore tune their systems with the training data. Then, in the test stage, the 10 chunks of the \mathcal{TE}_{DS} data 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 task were asked to give their predictions on the users contained in the \mathcal{TE}_{DS} , based on the partial information read from chunks ch_1 to ch_i .

2.2 Methods

In order to deal with the problem posed in the pilot task we proposed a new document representation named *temporal variation of terms* (TVT) [3]. TVT is an elaborate approach that requires enough space to be described in an adequate way. For this reason, in the present work we will only focus on the decision aspects that were considered at the pilot task to combine TVT with other well-known document representation approaches like *Bag of Words* (BoW). Thus, we will give below a separate description of the method with its main characteristics but the interest reader can obtain a detailed description of TVT in [3].

We used in our experiments different document representations and learning algorithms. Therefore, we start giving some general ideas of those document representations and a little more detailed explanation of TVT. After that, some general considerations on the used learning algorithms are briefly introduced.

Document representations

⁴ http://early.irlab.org/task.html

- Bag of Words. The traditional Bag of Words (BoW) representation is one of the language models most used in text categorization tasks. In BoW representations, features are words and documents are simply treated as collections of unordered words. Formally, a document d is represented by the vector of weights $d_{BoW} = (w_1, w_2, ..., w_n)$ where n is the size of the vocabulary of the dataset. Each weight w_i is a value that is assigned to each feature (word) according to whether the word appears in a document or how frequently this word appears. This popular representation is simple to implement, fast to obtain and can be used under different weighting schemes (boolean, term frequency (tf) or term frequency - inverse document frequency (tf - idf)). In our study we used BoW with the boolean weighting scheme.
- Concise Semantic Analysis (or Second Order Attributes (SOA)). Concise Semantic Analysis 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 [6] and has been adapted to work in author profiling tasks under the name of Second Order Attributes (SOA) [7].
- Character 3-grams. A n-gram is a sequence of n characters obtained from each text in the dataset. In the n-gram model, the dictionary contains all n-grams that occur in any term in the vocabulary. The representations using character n-grams have demonstrated to be effective in many applications [5] where n-grams are considered the terms used in BoW representations.
- LIWC-based features. Features derived from Linguistic Inquiry and Word Count (LIWC)[11, 10] have been used in several studies related to psychological aspects of individuals. LIWC has been successfully used to identify depressed and non-depressed people analyzing linguistic markers of depression such as the use of the personal pronouns and positive-negative emotions [12] and the presence of words related to the death (e.g., "dead", "kill", "suicide"), sex (e.g., "arouse", "makeout", "orgasm") and ingestion (e.g., "chew", "drink", "hunger") besides emotions also resulting useful [1]. Studies on suicidal individuals have incorporated LIWC as a valuable tool to extract information related to suicide and suicidal ideation analyzing the categories death, health, sad, they, I, sexual, filler, swear, anger, and negative emotions [2, 4]. Due to the fact we wanted to consider more meaningful features, we also considered in preliminary studies the most informative features belonging to linguistic dimensions (for example, personal pronouns and number of function words), summary of language variables (for example, dictionary words and words with more than 6 letters), psychological process (for example, negative emotions and affective processes) and, grammar (verbs and numbers) and punctuation (number of apostrophes) aspects.

Temporal Variation of Terms (TVT) The temporal variation of terms (TVT) method [3] is an approach for early risk detection based on using the variation of vocabulary along the different time steps as concept space for document representation. This method is the key component of our EDS system for the eRisk pilot task.

TVT is based on the concise semantic analysis (CSA) technique proposed in [6] and later extended in [7] for author profiling tasks. In this context, the underlying idea 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 chunks. These first chunks of the minority class, along with their complete documents, are 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 [3] showed its potential to deal with ERD problems.

Learning algorithms In preliminary studies we tested different learning algorithms and we obtained the best results with Random Forest and Naïve Bayes in several comparisons with other popular methods like LIBSVM⁵. We also used in our system, a decision tree (Weka's J48) obtained by first selecting the 100 words with the highest information gain and then removing from that list those words that were considered dependent of specific domains (names of politicians like "Obama" and countries like "China"). The J48 algorithm trained on this subset of features obtained a decision tree of only 39 nodes containing some interesting words like "meds", "depression", "therapist" and "cry", among others. As we will see later, this decision tree was used to assist to the TVT method in the initial chunks.

Even though we carried out several comparative studies with LIWC features, character 3-grams, CSA representations and the LIBSVM algorithm, we decided to select only those approaches that seemed to be effective to assist TVT in some specific situations. For this reason, in our ERD system we only used the Random Forest and Naïve Bayes algorithms with BoW and TVT representations and the J48 decision tree previously explained.

3 Training Stage

The pilot task was divided into a training stage and a test stage. So, in the first one, the participating teams had access to the \mathcal{TR}_{DS} set with all chunks of all training users. Therefore, we could tune in this stage our system with the training data.

⁵ For all the tested algorithms we used the implementations provided by the Weka data mining tool.

Even though the models obtained with the TVT representation showed in several preliminary studies to give better results than BoW and CSA (see [3] for a detailed study), we lacked of a robust criterion to solve a critical aspect in ERD problems that we will refer as the *classification time decision* (CTD) issue. That is, an ERD system needs to consider not only *which class* should be assigned to a document; it also needs to decide *when* to make that assignment. In other words, ERD methods must have a criterion to decide *when* (in what situations) the classification generated by the system is considered the final/definitive decision on the evaluated instances. Although this aspect has been addressed with very simple heuristic rules⁶, we wanted to have a rule for the CTD issue that attempted to exploit the strengths of the TVT method and simultaneously alleviate its weaknesses. With this goal in mind, we began a thorough study to determine in which situations the TVT method behaved satisfactorily and when it needed to be improved.

We started our study by dividing the training set \mathcal{TR}_{DS} into a new training set that we will refer as $\mathcal{TR}_{DS} - train$ and a test set named $\mathcal{TR}_{DS} - test$. Those sets maintained the same proportions of post per user and words per user as described in [8]. $\mathcal{TR}_{DS} - train$ 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 individuals (63 positive, 288 negative) meanwhile $\mathcal{TR}_{DS} - test$ contains 135 individuals (20 positive, 115 negative). The division into 10 chunks of the writings provided by the organizers was kept for both collections.

Then, we analyzed the performance of the TVT method by using the \mathcal{TR}_{DS} train set as training set and evaluating the obtained results with \mathcal{TR}_{DS} -test. We started 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 evaluated the TVT's performance considering that the model is (simultaneously) applied to the writings received up to the chunk \hat{C}_i . With this type of study it was possible to observe to what extent the TVT method was robust to the partial information in the different stages, in which moment it started to obtain acceptable results, and other interesting statistics. In this context, we considered that the F_1 measure which combines precision and recall would be a valuable measure to gain some insight into the performance of the TVT method along the different chunks.

Figure 1 shows this type of information where we can see that TVT representations with two different learning algorithms (Naïve Bayes (NB) and Random Forest (RF)) reached the highest F_1 values in the chunks 3 and 4. In both cases, those values were obtained when an instance was classified as positive if the probability assigned by the classifier was greater or equal than 0.6 ($p \ge 0.6$). Besides, we could observe that in the first chunk, TVT representations produced a high number of false positives. That weakness of TVT methods is shown in Figure 2 where we can see the lowest precision values in the chunk 1. That problem of TVT methods looks reasonable if we consider that TVT is based on the

⁶ For instance, exceeding a specific confidence threshold in the prediction of the classifier [8].

variations of terms along the time. Thus, we can conclude that in the first two chunks, where little information is available, TVT will need to be assisted by other complementary methods.



Fig. 1. F_1 values obtained with TVT models.

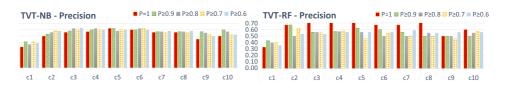


Fig. 2. Precision values obtained with TVT models.

Due to the fact that we wanted to focus our predictions on those chunks where best results were obtained (chunks 3 and 4) and assuming that after those chunks the penalization component in the *ERDE* computation would affect the TVT's results we decided to set some "chunk by chunk" rules that accomplish certain basic properties in the different chunks:

- 1. Chunk 1: Here the ERD system should be extremely conservative and only classifying an instance as positive ("depressed") if there exists strong evidence of that. We used for this case, the criterion of only classifying an instance as positive if *both models* obtained with TVT-NB and TVT-RF classified the instance as positive with probability p = 1 and the text included all the words in a "white list". That list was obtained from the words with the highest information gain of the documents of the first chunk. It included the words "depression" and "diagnosed".
- 2. Chunk 2: Here the restriction of the "white list" could be relaxed and complement the TVT methods with a more general approach. For this end, we used the predictions of the J48 decision tree explained above and classified an instance as positive if the three classifiers classified it in that way.
- 3. Chunk 3: In this chunk most of the classifiers obtained the best precision values. However, they obtained low recall values. In order to address this aspect, an instance was classified as positive if at least two classifiers classify it as positive with probability $p \geq 0.9$.

- 4. Chunk 4: BoW obtained in this chunk the highest recall values but low precision. We also could observe that when it was combined with the TVT method, this last method played a role of "filter" obtaining good results when both methods classified an instance as "positive". On the other hand, many instances not detected for this combination were classified as positive by the J48 method. For this reason, any instance classified as positive by both, BoW and TVT methods with probability $p \ge 0.7$ or by the J48 method was classified as positive.
- 5. Chunks 5 to 10: From chunk 5 forward, we assumed that most of the relevant classifications had already been made in the previous chunks. However, we kept a monitoring system to identify those cases that needed much more additional information to determine whether an individual was depressive or not. For this purpose, we used the same rule of chunk 4 for chunks 5 and 6 but increasing the probability of BoW and TVT to $p \ge 0.8$. From chunks 7 to 10 an instance was classified as positive if at least two classifiers classified it as positive with probability $p \ge 0.9$.

With those empirically derived rules as general criterion for the CTD issue, we then analyzed if the TVT method combined with the other approaches in specific situations obtained a better performance than the individual methods working separately. The results obtained from this analysis are presented below.

3.1 Analysis of results

In order to study the performance of our "combined" approach against each method tested separately, we used $\mathcal{TR}_{DS} - train$ and $\mathcal{TR}_{DS} - test$ sets in the experiments. Individual methods also need a criterion to deal with the CTD issue. In this case, we can 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 [8] is referred as dynamic, only considers that this probability exceeds some particular threshold to classify the instance/individual as positive. In our studies, we also used, for the individual methods, this dynamic approach by considering different probability values: p = 1, $p \geq 0.9$, $p \geq 0.8$, $p \geq 0.7$ and $p \geq 0.6$.

Tables 1, 2 and 3 report the values of precision (π) , recall (ρ) and F_1 measure (F_1) of the target ("depressed") class for each considered individual model and different probabilities. Statistics also include the *early risk detection error* (ERDE) measure proposed in [8] with two values of the parameter o used in the pilot task: o = 5 (*ERDE*₅) and o = 50 (*ERDE*₅₀).

As we can see from Table 1 the model TVT-NB obtained the best performance when the classifier assigned to the positive class (depressed), the instances with probability 1. For TVT-RF and BoW the statistics are different. In Table 2 the best $ERDE_{50}$ (and F_1 measure) was obtained when Random Forest classifier used the lowest probability ($p \ge 0.6$) but at the same time with this probability the $ERDE_5$ error is the worst value. The BoW model (Table 3) obtained the best $ERDE_{50}$ value considering $p \ge 0.8$ and with this configuration the best F_1 ,

 π and ρ are obtained. The J48 model assigns the instances to the positive class using probability 1, then its performance is directly shown in the comparison in Table 4.

	p = 1	$p \ge 0.9$	$p \ge 0.8$	$p \ge 0.7$	$p \ge 0.6$
$ERDE_5$	14.13	16.66	16.68	16.82	16.85
$ERDE_{50}$	11.25	15.34	15.21	15.24	15.07
F_1	0.40	0.27	0.28	0.27	0.28
π	0.47	0.50	0.48	0.43	0.44
ρ	0.35	0.18	0.19	0.19	0.20

Table 1. Performance of TVT-NB model ($\mathcal{TR}_{DS} - test$ set).

Table 2. Performance of TVT-RF model (\mathcal{TR}_{DS} - test set).

$p = 1 \ p \ge 0.9 \ p \ge 0.8 \ p \ge 0.7 \ p \ge 0.6$							
$ERDE_5$	16.92	16.85	16.92	16.96	17.14		
$ERDE_{50}$	16.43	16.05	16.12	16.02	15.79		
F_1	0.11	0.15	0.16	0.21	0.22		
π	0.50	0.54	0.50	0.50	0.43		
ρ	0.06	0.08	0.10	0.13	0.14		

Table 3. Performance of BoW model $(\mathcal{TR}_{DS} - test \text{ set}).$

	p = 1	$p \ge 0.9$	$p \ge 0.8$	$p \ge 0.7$	$p \ge 0.6$
$ERDE_5$	20.79	20.83	21.05	21.16	21.27
$ERDE_{50}$	18.82	18.66	18.13	18.24	18.35
F_1	0.19	0.21	0.24	0.24	0.23
π	0.11	0.13	0.14	0.14	0.14
ρ	0.5	0.65	0.75	0.75	0.75

To choice a particular probability to compare the performance of the models, we selected the one for which the model obtained the best ERDE considering o = 50 ($ERDE_{50}$). We decided to use this metric because we knew in advance that the results in the pilot task would be evaluated with ERDE but we did not know the specific o value. In this context, we considered that $ERDE_{50}$ would be more important/informative than the $ERDE_5$ error.

Table 4 shows the comparison of performance of all models. As we can observe, the performance of the combined methods was the best in all metrics except for the $ERDE_5$ error for which TVT-NB model obtained the best value. It is interesting to note that beyond that TVT-NB obtained a good $ERDE_5$ value, in the other metrics it was slightly worse than the combined methods. These results confirm the adequateness of the combined approach for the ERD task over the other individual methods.

Table 4. Performance comparison of all models $(\mathcal{TR}_{DS} - test \text{ set}).$

	$ERDE_5$	$ERDE_{50}$	F_1	π	ρ
Combined methods	15.36	9.16	0.59	0.48	0.75
J48	17.01	12.94	0.4	0.33	0.5
BoW $(p \ge 0.8)$	21.05	18.13	0.24	0.14	0.75
TVT-NB $(p=1)$	14.13	11.25	0.4	0.47	0.35
TVT-RF $(p \ge 0.6)$	17.14	15.79	0.22	0.43	0.14

4 Test 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 set. In this subsection, we show the results obtained by the combined methods trained with the full training set of the pilot task (\mathcal{TR}_{DS}) and tested with \mathcal{TE}_{DS} that was incrementally released during the testing phase of the pilot task.

In Table 5 we show the three submissions that obtained the best $ERDE_5$, $ERDE_{50}$ and F_1 values in the eRisk pilot task as reported in[9]. There, we can observe that our combined methods (UNSLA) obtained the best $ERDE_{50}$ value. On the other hand, FHDO-BCSGA obtained the best F-measure with a $ERDE_{50}$ value slightly worse than our approach. FHDO-BCSGB obtained the best $ERDE_5$ and the worst F_1 over the three approaches. With these results, we can conclude that our proposal is a competitive approach for ERD tasks.

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

	$ERDE_5$	$ERDE_{50}$	F_1	π	ρ
Combined methods (UNSLA)	13.66	9.68	0.59	0.48	0.79
FHDO-BCSGA	12.82	9.69	0.64	0.61	0.67
FHDO-BCSGB	12.70	10.39	0.55	0.69	0.46

5 Complementary studies on the test set

Our ERD system tested on the pilot task was derived from our analysis of the weakness and strengths of the TVT method on the training data. However, once the "golden-truth" information of the \mathcal{TE}_{DS} set was made available by the organizers, we could analyze what would have been the performance of the models, in particular the TVT method, working alone if different probabilities had been selected.

Table 6 shows this type of information by reporting the results obtained with TVT representation and using Naïve Bayes and Random Forest as learning algorithms. Besides, different probability values were tested for the dynamic approaches to the CTD aspect. The obtained results are conclusive in this case. TVT shows a high robustness in the *ERDE* measures *independently* of the algorithm used to learn the model, and the probability threshold. Most of the TVT's *ERDE*₅ values were low and in 7 out of 10 settings, the *ERDE*₅₀ values were lowest than the best one reported in the pilot task (the combined methods (*UNSLA*) with 9.68). In this context, TVT achieves the best *ERDE*₅ value reported up to now (12.30) with the setting TVT-RF ($p \ge 0.8$) and the lowest *ERDE*₅₀ value (8.17) with the model TVT-NB ($p \ge 0.8$). The performance of the TVT methods on the test set was surprising for us because on the training stage we could not obtain such as good results. This makes us conclude that if the TVT method had participated alone in the pilot task, it had obtained similar or better results than the ones obtained with the combined methods.

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

Table 6. TVT's performance on the \mathcal{TE}_{DS} set.

6 Conclusions and future work

This article presents the participation of LIDIC - UNSL at eRisk 2017 Pilot task on Early Detection of Depression. We proposed an interesting representation named *temporal variation of terms* (TVT) which considers the variation of

vocabulary along the different time steps as concept space for the representation of the documents. Because the experiments on the training stage showed some weakness of TVT, we proposed a combined approach to assist TVT in some specific situations. The results obtained on the pilot task with our combined methods were good, obtaining the best $ERDE_{50}$ value over all participants. An interesting aspect was that the TVT representation alone on the test set, obtained better results than the ones achieved by the combined approach.

As future work we plan to extend the analysis of TVT representation to solve other ERD problems such as the identification of sexual predators and people with suicide tendency.

References

- M. R. Capecelatro, M. D. Sacchet, P. F. Hitchcock, S. M. Miller, and W. B. Britton. Major depression duration reduces appetitive word use: An elaborated verbal recall of emotional photographs. *Journal of Psychiatric Research*, 47(6):809 – 815, 2013.
- M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pages 2098-2110, New York, NY, USA, 2016. ACM.
- M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. Garciarena Ucelay, and L. C. Cagnina. Temporal Variation of Terms as concept space for early risk prediction. In G. J. F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction.*, volume 10456. Springer, 2017.
- A. Garcia-Caballero, J. Jiménez, M. Fernandez-Cabana, and I. García-Lado. Last words: an liwc analysis of suicide notes from spain. *European Psych.*, 27:1 -, 2012.
- M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1):9-26, 2009.
- Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li. Fast text categorization using concise semantic analysis. *Pattern Recogn. Lett.*, 32(3):441–448, February 2011.
- A. P. López-Monroy, M. Montes y Gómez, H. J. Escalante, L. Villaseñor-Pineda, and Efstathios Stamatatos. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 89:134 – 147, 2015.
- 8. D. E. Losada and F. Crestani. A Test Collection for Research on Depression and Language Use, pages 28-39. Springer International Publishing, Cham, 2016.
- D. E. Losada, F. Crestani, and J. Parapar. eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In *Proceedings Conference* and Labs of the Evaluation Forum CLEF 2017, Dublin, Ireland, 2017.
- 10. J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. 2015.
- 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.
- 12. N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In *Proc. ICWSM 2008*, 2008.