k-TVT: a flexible and effective method for early depression detection

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

[1] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Universidad Nacional de San Luis (UNSL)
Ejército de los Andes 950, (5700) San Luis, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
{lcagnina, merreca, mjgarciarenaucelay, funezdario, villegasmariapaula74}@gmail.com

Abstract. The increasing use of social media allows the extraction of valuable information to early prevent some risks. Such is the case of the use of blogs to early detect people with signs of depression. In order to address this problem, we describe *k*-temporal variation of terms (*k*-TVT), a method which uses the variation of vocabulary along the different time steps as concept space to represent the documents. An interesting particularity of this approach is the possibility of setting a parameter (the *k* value) depending on the urgency (earliness) level required to detect the risky (depressed) cases. Results on the early detection of depression data set from eRisk 2017 seem to confirm the robustness of *k*-TVT for different urgency levels using SVM as classifier. Besides, some recent results on an extension of this collection would confirm the effectiveness of *k*-TVT as one of the state-of-the-art methods for early depression detection.

Keywords: Early Risk Prediction, Early Depression Detection, Text Representation, Semantic Analysis Techniques, Temporal Variation of Terms.

1 Introduction

Early risk detection (ERD) on the Internet is an important research area due to the impact it might have in fields like health when people suffer depression, anorexia or other disorders that can threaten life. In the field of security, there is a latent risk when criminals and sex offenders try to attack using web technologies. In this context, detection of depression is a major public health concern and a leading cause of disability. It is clear that the development of computational tools that help detecting depressed people, whether they are diagnosed as depressive or not, is becoming a very relevant task as demonstrated by the increase in publications in the specialized literature [1-4]. In particular, a scenario that has started to receive more attention is the one referred as early depression detection (EDD), that is, detecting depressive persons *as soon and accurate as possible* [5]. However, EDD, like other ERD tasks, usually presents several challenging aspects to the standard machine learning field: 1)

unbalanced data sets, 2) *classification with partial information* and, 3) *the classification time decision*.

An additional difficulty in research on EDD is the scarcity of resources (data sets) publicly available for experimentation and development of automatic detection systems. For this reason, the primary objective in [6] was to provide the first collection to study the relationship between depression and language usage by means of machine learning techniques. Another important contribution of that work is the proposal of a new error measure called Early Risk Detection Error ($ERDE_\sigma$) that simultaneously evaluates the accuracy of the classifiers and the delay in making a prediction. The σ parameter serves as the deadline for decision making, i.e. if a correct positive decision is made in time $k > \sigma$ it will be evaluated by $ERDE_\sigma$ as if it were incorrect (false positive). In that way, σ parameter allows to specify the urgency (earliness) level required for a task, that is to say, the lower the σ value the sooner a depressed user needs to be detected.

Beyond the effectiveness of TVT [7] in the first EDD pilot task, we found several limitations. First of all, the "heuristic" value of 4 "chunks" (short pieces of text) was completely empiric, that is, it mainly addressed the problem of balancing the minority class. However, no effort was dedicated to analyze what would be the impact of varying the number of chunks or determining what is the relation between the used number of chunks and the urgency level specified by the σ value. Besides, although different algorithms (like SVM, Random Forest, Multinomial Naive Bayes) were tested and different parameters were considered (like the probability threshold τ), no guidance was provided to select an adequate TVT configuration that produces acceptable results for the EDD task with different requirements in the earliness level.

The present work addresses the above mentioned drawbacks by presenting *k*-TVT, a generalization of the TVT method that allows to vary the number of chunks *k* considered for the minority class. Depending on the urgency (earliness) required in a particular scenario (specified by the σ parameter), it is possible to select for *k*-TVT a proper number of chunks *k* that obtains acceptable $ERDE_\sigma$ values. Furthermore, we provide some guidance about which could be robust learning algorithms and appropriate thresholds τ to be combined with the specific selected *k* value.

In order to fulfill these objectives, our study should be able to answer the following research questions:

- RQ1: Are the number of chunks *k* and the level of urgency σ related? If this is the case, which would be the correct values of *k* for the different σ values?
- RQ2: Once the number of chunks *k* has been selected, what is the influence of the probability threshold τ used to classify a user as depressed? Does it vary depending on the degree of earliness determined by σ?
- RQ3: Are the results obtained with those configurations of *k*-TVT comparable with other state-of-the-art methods?

Hence, in Section 2 we describe *k*-TVT, the method proposed in this article for EDD. Section 3 describes the data set used in our experiments. In Section 4, different studies are carried out to answer the research questions mentioned above. Finally, Section 5 summarizes the main conclusions obtained and possible future works are suggested.

2 The Proposed Method

Our method is based on the *concise semantic analysis* (CSA) technique proposed in [8] and later extended in [9] for author profiling tasks. Therefore, we first present in Section 2.1 the key aspects of CSA and then explain in Section 2.2 how we instantiate CSA with concepts derived from the terms used in the temporal chunks analyzed 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) [10] and *explicit semantic analysis* (ESA) [11] 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 $D = \{\langle d_1, y_1\rangle,\ldots,\langle d_n, y_n\rangle\}$ 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, $y_i \in C$ where $C = \{c_1,\ldots,c_q\}$ 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 $V = \{t_1,\ldots,t_v\}$ to the vocabulary of terms of the collection being analyzed.

2.1.1. Representing Terms in the Concept Space

In CSA, each term $t_i \in V$ is represented as a vector $t_i \in \mathbb{R}^q$, $t_i = \langle t_{i,1},\ldots,t_{i,q}\rangle$. 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 ith term and the jth concept, denoted w_{ij}, will be obtained. If $D_{c_u} \subseteq D$, $D_{c_u} = \{d_r \mid \langle d_r, y_s\rangle \in D \wedge y_s = c_u\}$ is the subset of the training instances whose label is the concept c_u, then w_{ij} might be defined as Equation 1 shows.

$$w_{ij} = \sum_{\forall d_m \in D_{c_j}} \log_2\left(1 + \frac{tf_{im}}{len(d_m)}\right) \quad (1)$$

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

As noted in [8] and [9], direct use of w_{ij} to represent terms in the vector $\boldsymbol{t_i}$ 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 [9]:

$$t'_{ij} = \frac{w_{ij}}{\sum_{i=1}^{|V|} w_{ij}} \quad (2) \qquad t_{ij} = \frac{t'_{ij}}{\sum_{j=1}^{|C|} w_{ij}} \quad (3)$$

where Equation 2 normalizes weights in proportion to the $|V|$ terms of each class and Equation 3 normalizes term weights in order to make them comparable among the $|C| = q$ categories/concepts. With this last conversion we finally obtain, for each term $t_i \in V$, a q-dimensional vector t_i, $t_i = \langle t_{i,1}, \ldots, t_{i,q} \rangle$ 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.

2.1.2. 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 [8]. Terms have distinct 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. A previous work in BoW ([12]) has 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 [9] for author profiling that represents each document d_m as the weighted aggregation of the representations (vectors) of terms that it contains (see Equation 4).

$$\boldsymbol{d_m} = \sum_{t_i \in d_m} \left(\frac{tf_{im}}{len(d_m)} \times \boldsymbol{t_i} \right) \quad (4)$$

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

2.2 *k*-Temporal Variation of Terms

In Subsection 2.1 we said that the concept space C usually corresponds to standard category names used to label the training instances in supervised text categorization tasks. In this scenario, which in [8] is referred as direct derivation, each category label simply corresponds to a concept. However, in [8] 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 [9] those ideas are extended by first clustering each category of the corpora and then using those subgroups (sub-clusters) as new concept space.[1]

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 [6] for research in ERD systems for depression cases.

Following [6], 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 one): $D_{I_l,1}, D_{I_l,2}, \ldots, D_{I_l,n_l}$. 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 has 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 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 $\hat{C}_m$ can be partitioned in two subsets $\hat{C}_m^+$ and $\hat{C}_m^-$, $\hat{C}_m = \hat{C}_m^+ \cup \hat{C}_m^-$ where $\hat{C}_m^+$ contains the positive cases of chunk $\hat{C}_m$ y $\hat{C}_m^-$ 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 ith chunk to the jth chunk (inclusive). Thus, $\hat{C}_{1-h}$ will represent the data set with the complete stream 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 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. That is, the number of documents of the majority/negative class ("non-depression") is significantly larger than of the minority/positive class (depression). More formally, following the previously specified notation $\left|\hat{C}_{1-h}^-\right| \gg \left|\hat{C}_{1-h}^+\right|$.

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

An alternative to balance the classes 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 *k-temporal variation of terms* (*k*-TVT) arises, which consists in enriching the documents of the minority class with the partial documents read in the first k 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 *k*-TVT we first determine the number k of initial chunks that will be used to enrich the minority (positive) class. Then, we use the document sets $\hat{C}_1^+$, $\hat{C}_{1-2}^+$, …, $\hat{C}_{1-k}^+$ 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 $(k + 2)$-dimensional space using the CSA approach explained in Section 2.1.

3 Data Set and Pilot Task

Our study was carried out on the data sets provided by *eRisk 2017 pilot task*[2] and *eRisk 2018 Lab*[3] on *early risk prediction* for *depression* [5, 13]. They are collections of writings (posts) of Social Media users taken from *Reddit*. There are two categories of users: "depressed" (or positive) and "non-depressed/control" (or negative). For each user, the data set contains a sequence of writings (in chronological order) 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. Table 1 summarizes, for both data sets, the number of users for each class.

Table 1. Data sets for depression task.

	Training		Test	
	Depressed	Non-Depressed	Depressed	Non-Depressed
eRisk 2017	83	403	52	349
eRisk 2018[4]	135	752	79	741

4 Experimental Study

We will use the data sets described previously: eRisk2017 and eRisk2018. Research questions RQ1 and RQ2, will be analyzed in Subsection 4.1. We use a cross-validation

[2] http://early.irlab.org/2017/index.html
[3] http://early.irlab.org/2018/index.html
[4] The eRisk 2018 training set is the join of eRisk 2017 training and test sets.

study on the whole data set eRisk2017 in order to evaluate if it is possible to select an appropriate number of chunks *k* to obtain good $ERDE_\sigma$. Also, the incidence of the probability threshold τ on the classifier's performance is deeply analyzed. In Subsection 4.2, experimental results are compared with the ones previously published in [5, 13]. Thus, research question RQ3 will be answered.

4.1 Setting the *k* Parameter

We present an exploratory analysis that allows us a deeper understanding of the relationship between the number of initial chunks *k* used by *k*-TVT and the urgency level specified by σ. In that way, we could give some guidance about a reasonable number *k* for the different thresholds σ.

To carry out this study we perform a 5-fold cross validation with different versions of *k*-TVT using 3 learning algorithms: Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest. The implementations of these algorithms correspond to those provided in the Python scikit-learn library with the default parameters.

The performance of the classifiers was assessed using the $ERDE_\sigma$ measure and the parameter σ was varied considering the values: 5, 10, 25, 50 and 75. Because σ represents some type of urgency in detecting depression cases, we want to analyze how *k*-TVT performs under different levels of urgency. Note that σ = 5 means a high urgency (a quick decision should be made) and σ = 75 represents the lowest urgency (there is more time to make a decision) in detecting the positive cases.

As we stated before, *k*-TVT defines concepts that capture the sequential aspects of the ERD problems and the variations of vocabulary observed in the distinct stages of the writings. Thus, different number *k* of chunks that will enrich the minority (positive) class could have an impact in the $ERDE_\sigma$ measure. In this study, the *k* value was varied in the (integer) range [0, 5].

In each chunk, classifiers usually produce their predictions with some "confidence", in general, the estimated probability of the predicted class. Therefore, we can select different thresholds τ considering that an instance is assigned to the target class when its associated probability *p* is greater (or equal) than certain threshold τ ($p \geq \tau$). Our study considered 4 different settings for the probabilities assigned for each classifier: $p \geq 0.9$, $p \geq 0.8$, $p \geq 0.7$ and $p \geq 0.6$. Note that once a classifier determines that an instance is positive in a specific chunk, that decision remains inalterable until chunk 10. Due to space constraints, only the best results are shown (Table 2).

The performance of *k*-TVT was compared against those obtained with a standard bag of words (BoW) representation with different weighting schemes: boolean, term-frequency (*tf*) and *tf*-inverse document frequency (*tfidf*). The best results with BoW were achieved with *tfidf* scheme, SVM as learning algorithm and different thresholds τ. These results were adopted as a baseline in the subsequent experiments.

Table 2 shows the best values obtained by *k*-TVT for the temporal-aware measure $ERDE_\sigma$ with the different urgency levels σ. As we can see in the first row of the table, for the lowest σ (σ = 5 and σ = 10), *k*-TVT obtains the best $ERDE_5$ and $ERDE_{10}$ (highlighted in boldface) with the minimum *k* value, that is, *k* = 0 and the SVM classifier using $p \geq 0.8$. Those two numbers are around a 5% better than the ones corresponding to the baseline (BoW-SVM, $p \geq 0.7$). However, it is interesting to notice in the same

row, the baseline is better than 0-TVT when σ = 25, 50 and 75 are used. Those results are a preliminary evidence that *k*-TVT performance, as measured by $ERDE_{\sigma}$, effectively depends on the selected number of chunks *k*. That is, 0-TVT does not seem to produce as good results for higher σ values as the ones obtained with σ = 5 and 10. The same fact is confirmed by the best $ERDE_{25}$ and $ERDE_{50}$ obtained by *k*-TVT using the first 4 initial chunks (*k* = 4), SVM as learning algorithm and a lower probability, *p* ≥ 0.7. Finally, the best $ERDE_{75}$ was obtained with 4-TVT using Naïve Bayes with the same probability (*p* ≥ 0.7). Here, the lowest $ERDE_{75}$ indicates that 4-TVT-NB outperformed the baseline in almost 1 unit which constitutes a good value.

Table 2. Best results of 5-fold cross validation on the whole eRisk2017 data set.

			ERDE				
		τ	5	10	25	50	75
Best $ERDE_{5-10}$	0-TVT-SVM	0.8	**13.58**	**12.48**	12.04	11.40	11.08
	BoW-SVM	0.7	14.13	13.18	11.75	10.97	10.49
Best $ERDE_{25-50}$	4-TVT-SVM	0.7	14.10	12.51	**11.00**	**9.57**	9.17
	BoW-SVM	0.6	14.42	13.20	11.28	10.38	9.70
Best $ERDE_{75}$	4-TVT-NB	0.7	14.49	12.72	11.05	9.67	**8.74**
	BoW-SVM	0.6	14.42	13.20	11.28	10.38	9.70

In summary, from this study we can conclude that when there is a high urgency level (low σ values) in detecting depression cases, the best performance is obtained with 0-TVT. As we decrease the level of urgency in the detection (σ ≥ 25), 4-TVT performs well and it can detect the positive cases with enough accuracy. It seems that while more information enriches the *k*-TVT representation, more confident can be the classifier, therefore better $ERDE_{\sigma}$ can be obtained. It is also worth to note that for σ ∈ {5, 10, 25, 50} SVM classifier obtained the best results demonstrating thus enough robustness. Using *k*-TVT the classifiers obtained the predictions with highest probability: 0.8 and 0.7 while if BoW is used, the threshold is lower (around 0.7 and 0.6).

Since SVM performs well in most of the cases, we can suggest it as an acceptable algorithm to be combined with generic *k*-TVT and it will be used in next subsection.

4.2 Performance of *k*-TVT - eRisk's Train and Testing Sets

Here we analyze our approach against some of the state-of-the-art methods in order to answer RQ3. In this way, our *k*-TVT results are directly compared with those obtained by the different groups participating in the tasks and published in [5,13]. Thus, we reproduce the same conditions faced by the participants: we first work on the data set released on the training stage for obtaining the models and then, these are tested on the test sets.

Table 3 shows the values obtained with *k*-TVT and SVM considering *p* ≥ 0.7 (4-TVT) and *p* ≥ 0.8 (0-TVT). The best values reported until now are also shown in this table. The complete description of those methods can be found in [5,13].

The results reveal several interesting aspects. First of all, we can confirm the hypothesis originated from the previous study regarding that lower *k* values for *k*-TVT

produce better $ERDE_\sigma$ when σ is low (high urgency level). Also, when there are low urgency levels, it is better to set *k* with higher values. With both *k*-TVT and probability thresholds (τ = 0.7 and 0.8), the $ERDE_5$ measures are better than the best published for the eRisk 2017 task. For $ERDE_{50}$ the 4-TVT outperforms the best published for eRisk 2017, while 4-TVT obtains a value slightly worse in eRisk 2018, although better than 0-TVT.

Table 3. Comparison between best results in eRisk 2017 pilot task and eRisk 2018 Lab.

eRisk 2017	$ERDE_5$	$ERDE_{50}$
0-TVT	**12.04**	10.67
4-TVT	12.66	**8.99**
FHDO	12.7	10.39
UNSLA[5]	13.66	9.68

eRisk 2018	$ERDE_5$	$ERDE_{50}$
0-TVT	**8.78**	7.39
4-TVT	8.82	6.95
FHDO-BCSGB	9.50	**6.44**
UNSLA[6]	**8.78**	7.39

5 Conclusions

In this article we present *k-temporal variation of terms* (*k*-TVT), a flexible and effective method for early depression detection. *k*-TVT considers the variation of vocabulary along the different time steps as concept space for document representation. The flexibility of *k*-TVT is given by the possibility of setting a parameter (the *k* value) depending on the urgency level (the threshold σ) required to detect the risky (depressed) cases.

We obtained interesting evidence about the relationship between the *k*-parameter and the required level of earliness (σ threshold) in the predictions. For low σ values (high urgency) a low number of chunks ($k = 0$) is an adequate representation while for low urgency (higher σ), the use of higher value ($k = 4$) seems to be better.

Interestingly 0-TVT and 4-TVT show to be competitive (in fact better) than state of the arts methods participating in the EDD tasks. Besides, a very relevant aspect of the *k*-TVT representation is the complete domain independence because it is only based on the vocabulary present in the training corpus. It does not depend on features specifically derived for the depression problem or other costly hand-crafted features. Even more, the mechanism used to determine the classification time does not need to be adapted to a particular domain. This makes *k*-TVT suitable for implementation in other early risk tasks such as the early detection of anorexia or pedophiles without virtually any cost of migrating from one domain to another.

In that context, as future work, we plan to apply the *k*-TVT approach to other problems that can be directly tackled as early risk detection such as sexual predation and suicide discourse identification.

[5] In eRisk 2017 pilot task UNSLA is an assembly of several methods which includes 4-TVT representation.

[6] In eRisk 2018 Lab, UNSLA is the same method 0-TVT-SVM.

Acknowledgments. This work was partially funded by CONICET and Universidad Nacional de San Luis (UNSL) - Argentina.

References

1. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., Eichstaedt, J. C.: Detecting depression and mental illness on social media: an integrative review. In: Current Opinion in Behavioral Sciences 18, pp 43-49. 2017.
2. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM'13), pp 128-137. 2013.
3. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, ACM, pp 47-56. 2013.
4. Park, M., McDonald, D. W., Cha, M.: Perception differences between the depressed and non-depressed users in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp 476-485. 2013.
5. Losada D.E., Crestani F., Parapar J. eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In: Jones G. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017. Lecture Notes in Computer Science, vol 10456, pp 346-360. Springer, Cham. 2017.
6. Losada, D. E., Crestani, F.: A test collection for research on depression and language use. In: Fuhr N. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2016. Lecture Notes in Computer Science, vol 9822, pp 28-39. Springer, Cham. 2016.
7. Errecalde, M. L., Villegas, M. P., Funez, D. G., Garciarena Ucelay, M. J., Cagnina, L. C.: 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.
8. Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K.: Fast text categorization using concise semantic analysis, Pattern Recognition Letters 32 (3), pp 441-48. 2011.
9. López-Monroy, A. P., Montes y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative subprofile specific representations for author profiling in social media. In: Knowledge-Based Systems 89, pp 134-147. 2015.
10. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. In: Journal of the ASIS 41 (6), pp 391-407. 1990.
11. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. In: JAIR 34 (1), pp 443-498. 2009.
12. Lan, M., Tan, C., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization, IEEE TPAMI 31 (4), pp 721-735. 2009.
13. Losada, D. E., Crestani F., Parapar J. Overview of eRisk: Early Risk Prediction on the Internet. In: Bellot P. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2018. Lecture Notes in Computer Science, vol 11018, pp 343-361. Springer, Cham. 2018.