UAM's participation at CLEF eRisk 2017 task: Towards modelling depressed bloggers

Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa, and Héctor Jiménez-Salazar

Language and Reasoning Research Group, Information Technologies Department,
Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa,
México City, México
{evillatoro,gramirez,hjimenez}@correo.cua.uam.mx

Abstract. In this paper we describe the participation of the Language and Reasoning Research Group of UAM Cuajimalpa at eRisk 2017 pilot task: Early Risk Prediction on the Internet. The goal of the eRisk task consists in detecting with enough anticipation cases of depression on texts. For evaluating this task, organizers provided a dataset containing comments from a set of Social Media users. All comments are chronologically ordered and represent writings from depressed and non-depressed users. Our proposed approach addressed this problem by means of graph models. This type of representation allows to capture some inherent characteristics from documents that can be determined though traditional graph measurements, and then, employed as features in a supervised classification system. Obtained results indicate that more experiments, as well as a more thorough analysis is required to conclude regarding the pertinence (or not) of our proposed strategy.

Keywords: N-gram graphs, Graph similarities, Early text classification, Natural Language Processing

1 Introduction

Social media is an excellent tool for anyone to express their opinion about any topic in any context; furthermore, social media is perhaps the most used communication channel nowadays. In spite of this easiness of communication, this type of tools comprise a major threat to users, who are exposed to a number of risks and potential attacks. Consider, for instance, the problem of detecting sexual predators approaching minors or the identification of aggressive users [1,2,3,4,5]. These threats pose a challenge to the research community, that has to develop protective and preventive tools for avoiding potential risks. As mentioned in [6,7], more recently, special attention has been paid to another type of threats, such as those menaces coming from the individuals themselves, for instance depression, mental situation that may lead to more complicated situations such as suicide.

Although considerable research has been devoted to detect different types of threats, most of the current solutions work in a forensic scenario, *i.e.*, they are applied once the treat or the attack has been accomplished. Even though these

forensic methods are very useful in some particular scenarios, it is known that preventive mechanisms would have a greater and immediate impact into users own security.

Accordingly, the eRisk [7] at CLEF-2017[1] proposes an exploratory task on early risk detection of depression. The challenge consists of sequentially processing pieces of evidence and detect early traces of depression as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media[2].

We address the eRiks problem as a Text Classification (TC) problem. However, contrary to most TC techniques, we go beyond the Bag-of-Terms (BoT) models, and instead we represent documents and categories as graph models. This type of representation is able to incorporate contextual information by means of considering *terms*[3] co-occurrence values, which is an important limitation of the traditional BoT representation [8]. Thus, our main hypothesis establishes that writing patterns, both from content and style, can be encapsulated by means of this type of representation and would provide to a classifier discriminatory information for learning to distinguish depressed bloggers.

The remainder of the paper is organized as follows. Section 2 describes the applied methodology; Section 3 shows our experimental setup and obtained results; finally, in Section 4 some conclusions of this work are presented.

2 Methodology

Our work was mainly inspired by the ideas proposed in [8,9]. Contrary to the traditional Bag-of-Terms representation model, a graph model considers the order of terms' appearances in the original text, thus incorporating valuable contextual information to the representation. Generally speaking, the graph model associates neighboring pairs of tokens with edges that denote their frequency of co-occurrence. As a result, documents with different term sequences end up having identical or at least highly similar representations. It is worth mentioning that authors from [8,9], performed and exhaustive evaluation of this approach on thematic text classification, but for the best of our knowledge, this approach has not been previously evaluated in a non-thematic text classification task. Accordingly, we consider as an important contribution of this paper the pertinence evaluation of this approach on the posed task.

Thus, for applying the method described in [8] we need to perform the following two steps: *i)* build the classes' graph, *i.e.*, a prototype graph model for each category; and, *ii)* extract similarity features to train a supervised classifier, each instance is represented by $4 \times N$ features (where N is equal to the number of categories, $|C|$).

As we mentioned, during the first step we create the prototype graph for each category $c_j \in C$. Thus, we represent each document d_i from the training set,

[1] Conference Labs of the Evaluation Forum: `http://clef2017.clef-initiative.eu`
[2] `http://early.irlab.org`
[3] Unless explicitly mention, a *term* can be either a character n-gram or word n-gram.

belonging to category c_j as the graph G^{ij}. Notice that each node from G^{ij} is a term (or token) w within the document d_i, and the edges of G^{ij} account for the number of co-occurrences of every pair of nodes (terms) in the same contextual window within d_i. For the experiments reported in this paper we employed as terms character n-grams of size 3, and single words; and the size of the contextual window was defined as a symmetric window of 4 terms, meaning two terms to the left and two terms to the right from token w.

Once the graph for every document d has been generated, its necessary to merge all the graphs from documents of the same class c_j, resulting in the prototype graph of the class. As established in [8], the prototype graph has the following properties: its edges (nodes) comprise the union of the edges (nodes) of the individual document graphs, and its weights are adjusted so that they converge to the mean value of the respective weights. At the end, the resulting prototype graph captures patterns common in the content and style of the entire category, such as recurring and neighboring character and word sequences.

For the second stage, the training phase, we need to compute the features for every new document in relation to the prototype graph models. Basically, we compute the similarity between documents and prototype graphs through the closeness of their respective graph representations. Similarly as the work described in [8] we employed the three measures proposed by the author and we incorporated a content based metric. Next we describe the intuitive ideas behind each metric, however for further reference please follow [8,9].

- *Containment Similarity* (CS). This metric expresses the proportion of edges of a graph G_k that are shared with a second graph G_l, *i.e.,* sequences of shared nodes and edges.

- *Value Similarity* (VS). This measure indicates how many of the edges contained in graph G_k are contained in graph G_l, as well, considering also the weights of the matching edges. Notice that VS converges to its maximum value for graphs that share both the edges and the corresponding weights.

- *Normalized Value Similarity* (NVS). This is a derived measure from the previous metric, but without considering the relative size of the graphs being compared.

- *Dice Similarity* (DS). It it worth mentioning that this metric was not originally considered in [8]. This metric accounts for the number of shared nodes between graphs.

As we mentioned before, every instance will be represented by a feature vector of size $4 \times N$; where N is the number of categories in the classification problem and 4 similarity measures, namely: CS, VS, NVS and DS.

3 Experiments and results

This section is organized as follows: first we describe the provided dataset for performing our experiments; next we provide details on every system's configuration; then a brief description of the evaluation metrics, and finally, we discuss the obtained results.

3.1 Dataset

The test collection for the pilot task was initially described in [6]. Consist in a collection of posts 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. In order to simulate an early detection scenario, the dataset was 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. In summary, the training set contained 486 users (83 depressed, 403 non-depressed) and the test set contained 401 users (52 depressed, 349 non-depressed). Further details can be found in [7].

3.2 Submitted runs

As we mentioned before, for generating the prototype graphs for each category (*depressed* and *non-depressed*), we followed the methodology described in section 2 using two different forms for defining the terms w: character 3-grams and single words. It is important to mention that our generated graph models were constructed using the 100% of the provided training data, *i.e.,* the prototype models do not consider the chronology of the writings. At this stage, we only wanted to evaluate if this type of representation was capable of modeling depressed blogs at any time. Next we briefly describe the general configuration of each of our competing system.

- **LyRA:** For this configuration, the prototype graph models are build using as w terms single words. The idea behind this approach was to evaluate if thematic correspondences (sequences of words) might be helpful in distinguishing the writing patterns of depressed users.

- **LyRB:** Contrary to the previous configuration, this systems employs character 3-grams as terms for the construction of the prototype graph models. The rationale idea behind using character n-grams is to represent content and style patterns in the graph model, characteristics that might result helpful when distinguishing among depressed users.

- **LyRC:** We refer to this configuration as an hybrid system given that each instance is represented by a feature vector of size $2 \times 4 \times N$, *i.e.,* 16 features extracted from LyRA and LyRB systems. With this setting we wanted to evaluate how complementary are both graph models (based on single words

and character n-grams), in solving the posed task.

- **LyRD:** This is referred as a conservative ensemble method. This configuration considers the output of the LyRA, LyRB and LyRC classification systems, and assigns the class *depressed* if and only if the three systems agree on assigning this category.
- **LyRE:** Similarly to the previous system, this is also an ensemble method but contrary to LyRD, this is a majority vote scheme among the decissions from LyRA, LyRB and LyRC. Another characteristic of this system is that it began to work until chunk 9, once enough information was accumulated in order to emit a more confident decision.

As described in [7], the test stage consisted of 10 sequential releases of data, and each participating system has to choose, for each user in the collection, between two options: *(a)* emitting a decision on the user (*i.e.* depressed or non-depressed), or *(b)* waiting to see more chunks. For all our proposed systems, if the classifier assigns the class "depressed" to an user in any chunk, we emitted the final decision, otherwise we choose the option "seeing more chunks". It is important to mention that all our experiments waited until the last chunk for emitting the "non-depressed" decision; this is, a user that is difficult to classify across chunks is at the very end (*i.e.*, chunk 10) acknowledged as non-depressed users. Finally, as our classification algorithm we employed a lazy method, namely k-nearest neighbors algorithm, particularly we employed the provided implementation by the Weka toolkit [10], with $k = 1$ using an Euclidean distance[4].

3.3 Evaluation

For reporting our obtained results we employ traditional set-based metrics such as F-measure, Precision and Recall. However, as described in [6], these metrics are time unaware, and for this reason we also report our results using the official ERDE metric. Intuitively, the ERDE metric considers the correctness of the (binary) decision and the delay taken by the system to make the decision.

3.4 Results

Figure 1 shows the behaviour of our proposed systems across different chunks. For this analysis we do not consider as final decision the classification results of preceding chunks, *i.e.*, we evaluate our classifiers performance at every chunk.

Several important aspects can be observed across systems, for instance, all configurations, except for LyRE, obtained good precision results between chunk 3 and chunk 6. In general, after chunk 5 (chunk 4 and 6 for the LyRC configuration), the performance decays very rapidly; however as more evidence (text from chunks) is obtained, the performance begins to recover up to chunk 10, which is an expected behaviour since our prototype graphs were built using the

[4] `http://www.cs.waikato.ac.nz/ml/weka/`

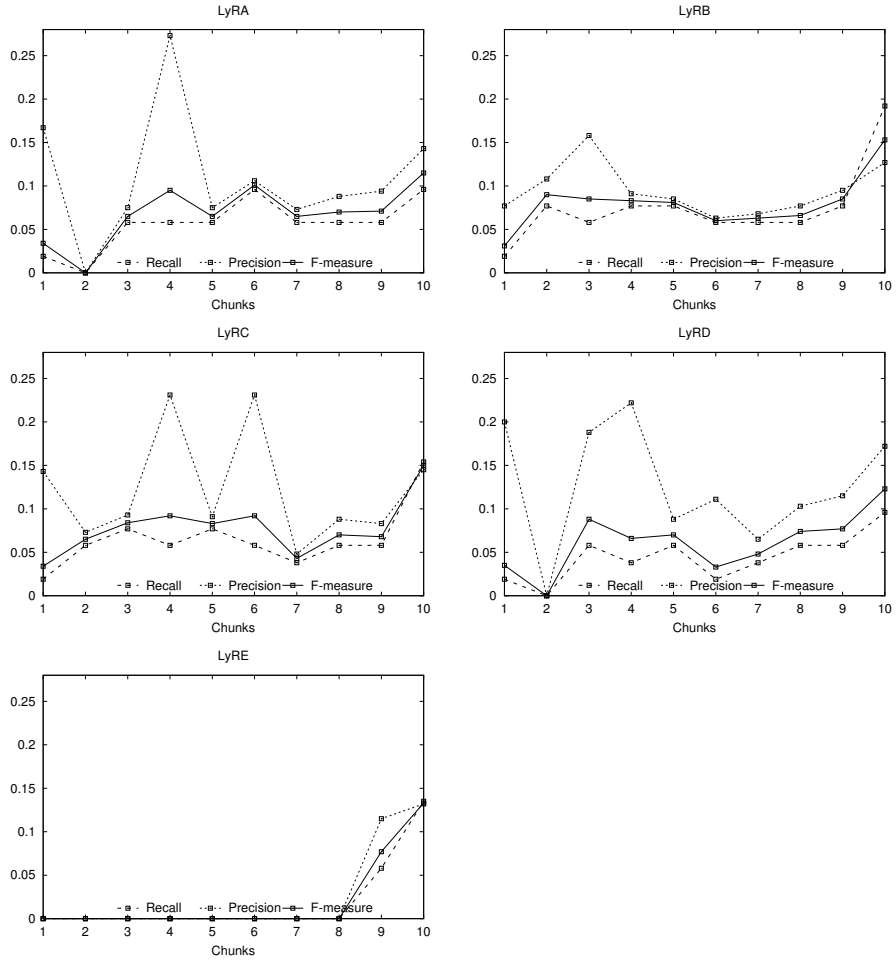

Fig. 1. Results for the positive class (*depressed*) of our proposed systems across the 10 chunks. We report Recall, Precision and F-measure obtained individually for every chunk.

Table 1. Official results during the eRisk competition. Last three rows represent the maximum (MAX), minimum (MIN) and Average values obtained across all participating systems.

System	$ERDE_5$	$ERDE_{50}$	F-measure	Precision	Recall
LyRA	15.65%	15.15%	14.00%	11.00%	19.00%
LyRB	16.75%	15.76%	16.00%	11.00%	29.00%
LyRC	16.14%	15.51%	16.00%	12.00%	25.00%
LyRD	14.97%	14.47%	15.00%	13.00%	17.00%
LyRE	13.74%	13.74%	8.00%	11.00%	6.00%
MAX	19.14%	17.15%	64.00%	69.00%	92.00%
MIN	12.70%	9.68%	8.00%	11.00%	6.00%
Average	14.67%	12.76%	39.77%	36.50%	51.17%

entire set of posts from depressed/non-depressed users. Although more experiments need to be done, our explanation for the peaks reached between chunks 3 and 6 is due to highly distinctive characteristics of depressed users at particular points of their writings. For proving this hypothesis, we need to test our proposed method considering the chronological order of the texts, in other words, to generate prototype graphs for different depression stages (chunks).

Regarding the *terms* considered by the graphs, our experiments indicate that both models, single words and character n-grams, are to some extent complementary to each other, e.g., observe that the LyRC is able to obtain in a couple of times high precision results. Similarly, the conservative ensemble (LyRD) shows a more stable performance (in terms of precision) in comparison to LyRA and LyRB configurations.

In Table 1 we observe the official results of our proposed systems during the eRisk competition. Notice that the last three rows from Table 1 report the maximum, minimum and average values for each metric obtained from all the participating systems[5]. According to the $ERDE$ metrics our best configuration was the LyRE system, however, as we mentioned before, this system began to emit decisions very late on the process, in chunk 9 (see Figure 1). For this reason, we assume as our best configuration the LyRD system, which obtains minimum error rates and better precision values.

4 Conclusions

In this paper, we have described the experiments performed by the Language and Reasoning Research Group from UAM Cuajimalpa in the context of the eRisk 2017 pilot task. Our proposed system was designed for addressing the posed problem by means of using as main form of representation graphs models; a graph model considers the order of terms' appearances in the original text, thus incorporating valuable contextual information to the representation.

[5] Further details on each participating system can be found in [7].

Even though this type of representation has been evaluated on thematic text classification, our main goal was to determine its pertinence on non-thematic classification tasks. Thus, the main hypothesis is that through this representation we can capture patterns common in the content and style from depressed and non-depressed users, such as recurring and neighboring character and word sequences.

From this exercise we have learned that our proposed method performs better when character n-grams are employed to build the graph models. However, combining this representation with graph models based on single words allows to identify more depressed users in early stages, but not in general. In addition, our results indicate that depressed users seem to reach a climax point during the first chunks of its writing, and as the times passes, they return to more "traditional" writing. Nonetheless, we need to perform more experiments in order to validate this hypothesis, for instance, to evaluate our proposal at modeling different depression stages.

As future work we plan to incorporate more features to our representation, for instance POS tags and word n-grams, which could help to identify more elaborated regularities among users. We also plan to model the writing characteristics from users at different chunks individually. And finally we want to perform experiments using different classification methods.

Acknowledgments. This work was partially funded by CONACYT under project grant number 258588. We also thank to UAM Cuajimalpa for the provided support.

References

1. G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at pan-2012.," in *CLEF (Online working notes/labs/workshop)*, vol. 30, 2012.
2. E. Villatoro-Tello, A. Juarez-Gonzalez, H. J. Escalante, M. Montes-y-Gomez, and L. Villaseñor-Pineda, "A two-step approach for effective detection of misbehaving users in chats," in *CLEF (Online Working Notes/Labs/Workshop)* (P. Forner, J. Karlgren, and C. Womser-Hacker, eds.), 2012.
3. H. J. Escalante, E. Villatoro-Tello, A. Juarez, M. Montes-y-Gomez, and L. Villaseñor, "Sexual predator detection in chats with chained classifiers," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, (Atlanta, Georgia), pp. 46–54, Association for Computational Linguistics, 2013.
4. Á. Callejas-Rodríguez, E. Villatoro-Tello, I. Meza, and G. Ramírez-de-la Rosa, *From Dialogue Corpora to Dialogue Systems: Generating a Chatbot with Teenager Personality for Preventing Cyber-Pedophilia*, pp. 531–539. Cham: Springer International Publishing, 2016.
5. K.-Y. Jeong and K.-S. Le, "Follower behavior analysis via influential transmitters on social issues in twitter," *Computacion y Sistemas*, vol. 20, no. 3, 2016.
6. D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Conference Labs of the Evaluation Forum*, p. 12, Springer, 2016.

7. 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.
8. G. Giannakopoulos, P. Mavridi, G. Paliouras, G. Papadakis, and K. Tserpes, "Representation models for text classification: A comparative analysis over three web document types," in *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, (New York, NY, USA), pp. 13:1–13:12, ACM, 2012.
9. G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos, "Summarization system evaluation revisited: N-gram graphs," *ACM Trans. Speech Lang. Process.*, vol. 5, pp. 5:1–5:39, Oct. 2008.
10. S. R. Garner, "Weka: The waikato environment for knowledge analysis," in *In Proc. of the New Zealand Computer Science Research Students Conference*, pp. 57–64, 1995.