LTL-INAOE's Participation at eRisk 2019: Detecting Anorexia in Social Media through Shared Personal Information

Rosa María Ortega-Mendoza, Delia Irazú Hernández Farías, and
Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
{rmortega, dirazuherfa, mmontesg}@inaoep.mx

Abstract. Detecting mental-health risk behaviours at primary stages is crucial to bring help to people and to avoid undesired consequences. In this paper, we describe our participation at the eRisk 2019 shared task. The proposed approach mainly relies on analysing the sentences that include personal information, i.e., fragments of texts reflecting user's interests, concerns, beliefs, personality traits and psychological state. The obtained results are very competitive, validating the important role of personal information for early detection of traces related to anorexia.

Keywords: Early Text Classification · Anorexia Detection · Personal Information.

1 Introduction

Many people around the world suffer some sort of mental health problem, but only a small portion of them receive treatment. Several problems that affect the society are strictly associated to mental illnesses, for example, violence, drug and alcohol abuse, among others. Therefore, taking care of mental health must be essential to warranty the well-being of society. There are some mental health issues which particularly affect to adolescents, such as *Anorexia* and *Self-harm*. Anorexia is an eating disorder that negatively impacts the relationship of the patient with food consumption. It is characterised by an uncontrolled concern about the weight, even refusing food. On the other hand, self-harm refers to a behaviour characterised by self-injury usually performed without suicidal intent. It is usually associated with the action of cutting, burning, scratching, or hitting body parts. Therefore, it is mandatory to dedicate efforts as a society to prevent and reduce the negative impact that such issues could cause.

Nowadays, people use social media as their main communication channel. Such platforms allow users to express their ideas, thoughts, and personal experiences. This makes user-generated content a very powerful source of data for research purposes. Recently, taking advantage of information coming from social media has attracted the attention of many studies to identify traces of problems related to mental health issues such as depression, anorexia, self-harm, etc. In the past few years, natural language processing methods have been used in order to develop automatic systems able to identify potential risks related to mental health on user-generated content. Such kind of research could serve as an additional resource for complementing traditional techniques used for analysing the people's behaviour from a social and psychological perspective.

From a computational linguistics point of view, it is possible to take advantage of the wide range of methods and algorithms in order to discover particular patterns depicted in a given text. Thus, helping to characterise text written by people having mental health issues. However, dealing with such a challenging task is strongly related to an important aspect: it needs to be successfully accomplished as soon as possible, i.e., the early detection of such risks. Most of the proposed approaches using user-generated content are designed for addressing such problems without paying enough attention to the importance of detecting potential risks in their first stages. Nevertheless, regarding the trace of mental health issues, the longer it goes without detection, the more likely is to increase a life-threatening. This has derived into the development of a challenging research area: the early risk (eRisk) prediction.

Since 2017 until now a shared task on eRisk prediction has been organised. This year, the *eRisk2019* was composed by three subtasks: (I) Early detection of signs of anorexia; (II) Early detection of signs of self-harm; and (III) Measuring the severity of the signs of depression. We participated only in the first two subtasks. For more details on the shared task see [9]. In order to address the Task I, we proposed an approach that mainly relies on the personal information reflected on the written texts. The main intuition is that such kind of text serves to capture the interest, habits, personality and psychological state of the users. For Task II, we developed a simple approach based on the similarity between a given piece of text and a set of phrases potentially related to self-harm. In order to calculate the similarity, we exploited the *Textual Attraction Force* (TAF) [1]. It is a novel paradigm for text classification that considers not only the distance between two documents but also their relevance in terms of a given criterion. In our case, we defined a vocabulary of self-harm related terms for calculating the relevance of each document. Besides, the distance between two instances was assessed by cosine similarity. Then, the presence of self-harm content was determined depending on the TAF between the document in hand and a set of instances including self-harm content. In this case, the organisers did not provide data for training purposes. Therefore, we retrieved data from Reddit by selecting a subset of posts published under the *"self-harm"* topic[1]. Unfortunately, the obtained results on Task II were not as expected. Further work needs to be done on this approach. For this reason, in this paper, we decided to focus our attention on the proposed method for detecting the presence of anorexia-related content (Task I).

[1] https://www.reddit.com/r/selfharm/

2 Related work

Early detection of risks related to mental health involves different disciplines such as psychology and linguistics. There is a strong relationship between the psychological state of the people with the language they use [6]. Social media content has been already exploited as a source of data for studying mental health-related issues. During the last few years, the CLPsych[2] workshop has been organised with the aim of promoting the interdisciplinary study of mental health care. It serves as a forum for sharing ideas from both psychological and computational linguistics perspectives. Different mental health-related status have been addressed by exploiting natural language processing approaches. There are some works in the literature focused on detecting depression in Twitter content [4, 15, 11]. A linguistic analysis for classifying patients suffering depression and paranoia was carried out by Oxman et al. [14]. With the aim to detect content revealing risk behaviours related to self-harm, Wang et al. [19] used data from Flickr, while Yates et al. [20] exploited posts coming from Reddit.

Detecting traces of anorexia is a challenging task that has been addressed from a psychological perspective [10, 5]. Recently, some approaches for dealing with such a task from a computational linguistics perspective have been proposed. A study on the content related to anorexia that is shared in Tumblr is described in [3]. Spinczyk et al. [17] analyzed a set of texts written by patients with anorexia and healthy people. In both works, the authors identified the use of words related to negative emotions in people having this eating disorder. Cavazos et al. [2] carried out an analysis of the content of a set of tweets containing a set of keywords related to eating disorders. Wang et al. [18] explored among eating-disorder communities on Twitter by sampling profiles of users self-identified with an eating disorder together with their social network connection.

Dealing with such challenging tasks has been considered as the main focus of some evaluation campaigns. The second edition of the early risk detection was organised in 2018 [8] covering two different tasks aimed to detect early risks of depression and anorexia. The main aim of the tasks was: given a sequence of writings in chronological order, to detect early traces of *depression/anorexia* as soon as possible. The participating systems proposed several approaches exploiting different machine learning algorithms as well as Convolutional Neural Networks. Several kinds of features were used such as the traditional bag-of-words, word embeddings, user-level meta-data, some based on semantic and psychological aspects, and domain-specific vocabularies. Most of the teams participated in both tasks using similar approaches, the main difference among them was the use of lexicons related to each of the tasks. The organizers of the task observed that most participating systems took a decision on the presence of *depression/anorexia* until hundreds of writings per user were processed. According to them, most of the teams concentrated their effort in terms of accuracy rather than taking decisions for avoiding delay.

In a previous work [12], the role of personal information for identifying personal traits (particularly age and gender) on social media texts was studied. Besides, in the 2018 edition of eRisk, an approach based on this kind of information was proposed [13]. The findings indicated that the most relevant features for distinguishing among

[2] http://clpsych.org/

profiles are comprised on personal phrases. Attempting to take advantage of such kind of data, we proposed an approach for early detection of anorexia by exploring the use of personal information.

3 Proposed methodology for anorexia detection

The *Task 1: Early Detection of Signs of Anorexia* has as the main goal to detect early traces of anorexia by processing sequentially pieces of Social Media texts of a given author. This task was organized also in 2018, however, this year a new schema for releasing the data was applied. Instead of made available the data exploiting a chunk-based approach, an item-by-item strategy was used. Two different stages were involved in the task: i) a training stage where a set of writings belonging to labelled users was released, and ii) a test stage, where iteratively user writings were provided, and then, before receiving a new fragment of data, the system must send a decision about each user.

Our participation in the eRisk 2019 challenge is based on the DPP-EXPEI approach. The approach was introduced and successfully exploited for author profiling [12]. Based on the idea that personal phrases help to highlight information that reveals the behaviour or mental state of people, this approach emphasises the value of the terms located in personal phrases. Specifically, we considered as personal phrases those containing the following pronouns: *I*, *me*, *mine*, *myself*, *my*, and *Im*. Following we describe the approach.

3.1 The DPP-EXPEI approach

The DPP-EXPEI approach serves to represent a given text by paying special attention to those terms that are located in personal phrases. It involves a two-stage process: First, it exploits a feature selection technique called *discriminative personal purity (DPP)*. Then, in order to assign a weight to each term, it uses a scheme denoted as *exponential reward of personal information (EXPEI)*. Below we briefly introduce both methods. For more details about DPP and EXPEI refer to [12].

Feature selection using DPP. Aiming to identify the most relevant terms for representing a text, DPP considers those terms that appear inside personal phrases. In order to determine which are the features to be selected, each term has associated a score according to both the overall distribution of all terms across the categories as well as the type of phrases it appears in. Formula 1 describes the DPP scheme; it considers the level of occurrence of the terms in personal phrases (their personal purity, PP) together with an estimation of the distribution of the terms across the categories. In the original scheme, the distribution was estimated by means of the *Gini* coefficient. In this work, we use the *dif* function as an effort to deal with the class imbalance and to emphasize the interest class (i.e., *depressive*).

$$DPP(t_i) = \max_{k=1}^{|C|} \{PP_k(t_i)\} \cdot dif(t_i), \tag{1}$$

where $dif(t_i)$ represents the difference on the number of documents containing the term t_i in the positive class (we applied a square root to this value to emphasize its importance) and the negative class.

Term weighting using EXPEI. When using a bag-of-words representation, the terms in a given document are scored by a weight aimed to determine the degree of contribution to describe the content of the document at hand. In particular, EXPEI considers the well-known normalized frequency (TF) weighting together with the occurrence of the terms in personal phrases according to the PEI value, as it is shown in Formula 2. The PEI measure estimates the quantity of personal information revealed by a term, rewarding the terms with a high concentration of personal information.

$$w_{ij} = \left(\sqrt{TF(t_i, d_j)}\right)^{1-PEI(t_i,d_j)} \tag{2}$$

where $TF(t_i, d_j)$ represents the normalized frequency of t_i in d_j.

3.2 Early risk detection based on DPP-EXPEI

For each user, his/her writings are represented with a vector. These vectors consider the 1,000 most discriminative terms according to the DPP values. The terms in the text representation are mainly lexical unigrams; the function words were excluded. These terms were weighted by the EXPEI scheme. For the classification phase, we used a linear Support Vector Machine (SVM) with L2 norm. Additionally, we designed some criteria to take a decision regarding early detection; these criteria are described below.

Criteria for early decision. In this work, the early detection is tackled by external criteria aimed to review the classifier decisions. We used two different criteria (C1 and C2) to decide whether to submit a decision for a subject or to wait for more writings:

- C1: to assign the positive decisions taken by the classifier in the current round.
- C2: to submit a positive decision only if the instance was classified as positive and it contains at least three terms from the top-50 terms selected by DPP.

In this edition of the task, the participating systems were asked to calculate a score estimating the level of anorexia of each user. We computed this score (sc) by counting the number of words with high DPP value in the users' posts. That is, for each writing (round number) r of a subject i, $sc_{i,r} = (f_{i,r} + 1)/(r + 1)$, where $f_{i,r}$ represents the occurrences of the n terms with the highest DPP values in all previously seen writings ($0...r$) from subject i. The score is estimated only for positive decisions, and it is re-computed in each writing round. In the experiments, we considered $n = 50$.

4 Experiments

4.1 Datasets

The datasets presented in the CLEF 2019 eRisk forum [9] consist of writings (posts or comments) from a set of social media users. There are two categories of users: with anorexia and non-anorexia. For each user, the collection contains a sequence of writings in chronological order. The training data corresponds to the collection of eRisk 2018 challenge, which is described in Table 1.

Table 1: Number of users by category in the training dataset

Subjects	Training	Test
with anorexia	20	41
non-anorexia	132	279
Total	152	320

4.2 Evaluation

The submitted runs were scored by some decision-based metrics. The purpose was to take into account not only the correctness of the decision by means of F_1 measure but also the delay by the approach to make the decision by metrics for early risk prediction. Specifically, to associate a cost to the delay in the detection of true positives, some measures were considered: the early risk error ($ERDE_o$) [7], the $latency_{TP}$, the speed and the latency-weighted F_1 [16]. These measures increase the cost according to the number of seen writings (items) for taking the decision. Particularly, $ERDE_o$ uses a sigmoid-like cost function and it was estimated with the cutoff parameter o set to 5 and 50 items (denoted as $ERDE_5$ and $ERDE_{50}$, respectively). $latency_{TP}$ assesses the system's delay based on the median number of observed (processed) writings to detect such positive cases. The system's overall speed factor qualifies the delay by means of a penalty factor. Finally, the *latency-weighted* F_1 combines the effectiveness of the decision (by the F_1 measure) and the delay.

Additionally, the evaluation was complemented applying some ranking-based metrics from IR. For each participating system, it was built a decreasing ranking of the users by their level of anorexia (or estimated risk) at each point. Each ranking was scored using the P@10 and NDCG metrics, which were applied after seeing k writings (k = 1,100,500,1000).

4.3 Results

We trained a model using the dataset of eRisk 2018. Specifically, we used the train partition of the dataset and we extended the anorexia evidence aggregating the whole writing histories of users with anorexia in the test partition of the same dataset[3]. This model was applied in the test stage of the competition. The obtained results are shown in Table 2[4]. It is possible to observe that both early designed criteria performed in

[3] We referred to the test partition of the eRisk 2018 dataset, therefore the ground truth labels are known.

[4] For the sake of the readability in this paper we used C1 and C2 to denote the submitted runs, while in the official results of the task our runs are denoted as "0" and "1", respectively.

a very similar way. We observed that the differences between both criteria are more evident in the first rounds, which have less information. In the subsequent rounds, the size of texts increased, augmenting the possibility of finding the terms with the greatest DPP values; in consequence, their differences are weakened. This suggests that a more accurate prediction could be obtained considering the number of occurrences of words with major DPP in the criterion C1, instead of only considering the number of words (without repetitions).

Table 2: Official results of the decision-based evaluation

run	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	*speed*	*latency-weighted F1*
C1	0.45	0.75	0.57	0.08	0.04	11	0.96	0.54
C2	0.47	0.75	0.58	0.08	0.04	11	0.96	0.55
avg	0.41	0.59	0.42	0.09	0.06	54.15	0.93	0.38

Regarding the official results, the proposed method demonstrated a competitive performance, since our results are considerably better than the average (avg) methods' performances in terms of each decision-based metric. For example, regarding $ERDE_5$ and $ERDE_{50}$ values, the method obtained very small differences against the best values (0.06 and 0.03 respectively). On the other hand, considering the F_1 measure, the runs of the proposed method were ranked among the top 14 and 15 positions. In this case, only the runs from five teams outperformed our results. A competitive performance was observed by the $latency_{TP}$ and $speed$ measures. On the other hand, considering the effectiveness of the decision and the delay by the latency-weighted F1 measure, the method was ranked among the top 12 and 15 positions. Similarly, the runs of four and five teams outperformed our results. These results show the usability of the DPP-EXPEI in anorexia detection.

Additionally, a ranking-based evaluation was implemented to assess the ranking of the users in accordance with their estimated level of anorexia. As previously explained, we designed a score based on a domain-specific vocabulary extracted by the DPP scheme. Examples of words in this vocabulary are: *anorexia, disorder, lbs, calories, anxiety, diet, stomach, weight, fat, therapy, foods, healthy, protein, snacks, gained, exercise, intake, gym*, among others. The performance of the approach using this score is shown in Table 3. The approach achieved results better than the average values of all runs in the competition, except to P@10 for 1000 writings where a small difference with the average was achieved. Note that, the proposed score was favoured when 100 and 500 writings were processed. We believe such behaviour is due to the fact that one writing has scarce personal information and 1000 writings have a high overlap with the defined domain vocabulary. However, it is important to highlight that these results show that the approach is totally reliable to detect the top-10 users with the highest risk.

Furthermore, the high values of P@10 and NDCG@10 indicate that most of the users qualified with high levels of anorexia (around 80%) were correctly classified even though considering one single writing. These results suggest that the proposed score is very informative, and confirms that words with high DPP values are highly relevant to identify anorexia.

Table 3: Official results of the ranking-based evaluation

	run	P@10	NDCG@10	NDCG@100
1 writing	C1	0.80	0.75	0.34
	C2	0.80	0.75	0.34
	avg	0.50	0.47	0.32
100 writings	C1	1.00	1.00	0.76
	C2	1.00	1.00	0.76
	avg	0.72	0.72	0.56
500 writings	C1	0.90	0.92	0.73
	C2	0.90	0.92	0.73
	avg	0.74	0.74	0.62
1000 writings	C1	0.70	0.78	0.65
	C2	0.70	0.78	0.66
	avg	0.72	0.72	0.62

5 Conclusions and future work

In this paper, we have described our participation at the 2019 CLEF eRisk Lab. We addressed the anorexia detection task applying the DPP-EXPEI approach. This is inspired by the relevance of personal information explored in some previous works. The main idea is that phrases with first-person pronouns contain information about users that can reveal mental health disorders such as anorexia. Consequently, the approach emphasizes the value of this information by means of term selection and weighting methods: DPP and EXPEI respectively.

The approach showed a competitive performance in the anorexia detection task according to the evaluation based on decision effectiveness and ranking. In general, the approach achieved higher results than the average results for each metric. The results evidenced the appropriateness of the DPP-EXPEI approach for the early risk detection of anorexia on social media texts. Besides, the results showed the suitability of the proposed score to estimate the level of anorexia. This score relates the level of anorexia of each user with the texts' concentration of a domain-specific vocabulary, which is formed by the terms with the highest DPP values. These findings allow to confirm that the personal information shared in social media concentrates a special value for detecting eating disorders. As future work, we are interested in deeply analyzing the relation of personal information with the anorexia disorder as well as with other risk states in mental health such as self-harm.

Acknowledgments

This research was funded by CONACYT (postdoctoral fellowship CVU-174410 and project FC 2016-2410).

References

1. Aguilera, J., González, L.C., Montes-y Gómez, M., Rosso, P.: A New Weighted k-Nearest Neighbor Algorithm Based on Newton's Gravitational Force. In: Vera-Rodriguez, R., Fierrez, J., Morales, A. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. pp. 305–313. Springer International Publishing, Cham (2019)
2. Cavazos-Rehg, P.A., Krauss, M.J., Costello, S.J., Kaiser, N., Cahn, E.S., Fitzsimmons-Craft, E.E., Wilfley, D.E.: "I just want to be skinny.": A Content Analysis of tweets Expressing Eating Disorder Symptoms. PLOS ONE **14**(1), 1–11 (01 2019)
3. De Choudhury, M.: Anorexia on tumblr: A characterization study. In: Proceedings of the 5th International Conference on Digital Health 2015. pp. 43–50. DH '15, ACM, New York, NY, USA (2015)
4. 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. pp. 47–56. WebSci '13, ACM, New York, NY, USA (2013)
5. Delinsky, S.: Body image and anorexia nervosa. Body Image, Second Edition: A Handbook of Science, Practice, and Prevention. (2nd ed.) pp. 279–287 (2011)
6. Holtgraves, T.: The Oxford Handbook of Language and Social Psychology. Oxford legal philosophy, Oxford University Press (2014)
7. Losada, D.E., Crestani, F.: A Test Collection for Research on Depression and Language Use. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 28–39. Springer International Publishing, Cham (2016)
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
9. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)
10. Lyons, E.J., Mehl, M.R., Pennebaker, J.W.: Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. Journal of Psychosomatic Research **60**(3), 253–256 (2006)
11. Nadeem, M.: Identifying depression on twitter. CoRR **abs/1607.07384** (2016)
12. Ortega-Mendoza, R.M., López-Monroy, A.P., Franco-Arcega, A., Montes-y-Gómez, M.: Emphasizing Personal Information for Author Profiling: New Approaches for Term Selection and Weighting. Knowledge-Based Systems **145**, 169 – 181 (2018)
13. Ortega-Mendoza, R.M., López-Monroy, A.P., Franco-Arcega, A., Montes-y-Gómez, M.: PEIMEX at erisk2018: Emphasizing personal information for depression and anorexia detection. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018)
14. Oxman, T., Rosenber, S., GJ, T.: The language of paranoia. American Journal Psychiatry **139**(3), 275–82 (1982)
15. Park, M., Cha, C., Cha, M.: Depressive Moods of Users Captured in Twitter. In: Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD) (2012)
16. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 495–503. WSDM '18, ACM, New York, NY, USA (2018)

17. Spinczyk, D., Nabrdalik, K., Rojewska, K.: Computer aided sentiment analysis of anorexia nervosa patients' vocabulary. BioMedical Engineering OnLine **17**(1), 19 (Feb 2018)
18. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 91–100. WSDM '17, ACM, New York, USA (2017)
19. Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., Chang, Y.: Understanding and discovering deliberate self-harm content in social media. In: Proceedings of the 26th International Conference on World Wide Web. pp. 93–102. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017)
20. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2968–2978. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)