Early Detection of Signs of Pathological Gambling, Self-Harm and Depression through Topic Extraction and Neural Networks

Diego Maupomé, Maxime D. Armstrong, Fanny Rancourt, Thomas Soulas and Marie - Jean Meurs

Université du Québec à Montréal, Montréal, QC, Canada

Abstract

The eRisk track at CLEF 2021 comprised tasks on the detection of problem gambling and self-harm, and the assessment of the symptoms of depression. RELAI participated in these tasks through the use of topic extraction algorithms and neural networks. These approaches achieved strong results in the ranking-based evaluation of the pathological gambling and self-harm tasks as well as in the depression symptomatology task.

Keywords

Mental Health, Natural Language Processing, Topic Modeling, Word Embeddings, Neural Networks, Nearest Neighbors

1. Introduction

This paper describes the participation of the RELAI team in the eRisk 2021 shared tasks. Since 2017, the eRisk shared tasks have aimed to invite innovation in Natural Language Processing and other artificial intelligence-based methods towards the assessment of possible risks to mental health based on online behavior [1]. In 2021, three tasks were put forth. The first task (Task 1) introduced the problem of detecting the signs of pathological gambling. The second task (Task 2), a continuation of Tasks 2 and 1 from 2019 and 2020, respectively, focuses on the assessment of the risk of self-harm. Finally, continuing from Tasks 3 and 2 from 2019 and 2020, Task 3 asked participants to predict the severity of depression symptoms as defined by a standard clinical questionnaire.

2. Task 1: Early Detection of Signs of Pathological Gambling

Pathological gambling is a public health issue with prevalence rate between 0.2% and 2.1% [2]. Accessing treatment is difficult since general practitioners usually do not screen for this pathology [3] and, by the time the issue has become evident, the patient has lost control [4].

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

maupome.diego@courrier.uqam.ca (D. Maupomé); darmstrong.maxime@courrier.uqam.ca (M. D. Armstrong); rancourt.fanny.2@courrier.uqam.ca (F. Rancourt); meurs.marie-jean@uqam.ca (M. J. Meurs)
0000-0001-8196-2153 (M. J. Meurs)

With the rise of online platforms, more data are available for potential detection systems [5]. Recently there has been research work focused on communications with customer services to detect whether a subject was at risk of gambling [6, 4]. However, to the best of our knowledge, textual productions from online fora have not yet been used to detect early signs of pathological gambling.

2.1. Task and Data

As mentioned, the data issue from Reddit users. These subjects have been labeled as either at risk for pathological gambling (positive) or not (negative). No labeled data were provided for training models. As such, the following pertains solely to the test data.

The test data comprised 2348 subjects, 164 of which were positive (6.9%). The test data are released iteratively, with each step counting at most one writing per subject. These writings are sorted in chronological order of publication. The first iteration includes writings from all test subjects. Thereafter, subjects are included as long as they have unseen writings.

Algorithms are expected to predict both a binary label and a score at each step. The label can default to negative. However, a positive prediction for a given user is binding, and all label predictions thereafter are disregarded. Evaluation, which is detailed in the following subsection, considers the labels and the timeliness of positive predictions, as well as the scores.

2.2. Evaluation

Performance is measured both on the ultimate decision made on each subject, using binary classification metrics, and the predicted scores, using ranking metrics. The classification metrics include standard precision, recall and the associated F_1 score, as well as Early Risk Detection Error ($ERDE$), $latency_{TP}$, $speed$ and $F_{latency}$. In addition to accounting for the binary prediction on a given subject, $ERDE$ seeks to account for the timeliness of that prediction by counting the number of writings processed by the predicting algorithm before producing a positive prediction. Given t_u, p_u, respectively the ground truth and predicted labels for a given subject, k_u the number of processed writings, and a given threshold o, the $ERDE$ is computed as:

$$ERDE_o(p_u, t_u, k) = \begin{cases} c_{fp} & \text{if } p_u \neq t_u = 0 \\ 1 & \text{if } p_u \neq t_u = 1 \\ \frac{1}{1+e^{o-k}} & \text{if } p_u = t_u = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, c_{fp} is a constant set to the rate of positive subjects in the test set. This per-subject $ERDE$ is averaged across the test set, and is to be minimized. Thus, false negatives are counted as errors, and false positives are counted as a fraction of an error in proportion to the number of positive subjects. The delay in decision is only considered for true positive prediction, where a standard sigmoid function counts the number of writings processed, k, offset by the chosen threshold, o. As with the other classification metrics used, true negatives are disregarded. $ERDE$ was evaluated at $o = 5$ and $o = 50$.

Likewise, $latency_{TP}$ measures the median delay in true positive predictions:

$$latency_{TP} = median\{k_u : u \in U, p_u = t_u = 1\}$$

Similarly, $speed$ and $F_{latency}$ [7] are computed by penalizing this delay albeit in a smoother manner:

$$speed = 1 - \text{median}\left\{penalty(k_u) : u \in U, p_u = t_u = 1\right\}$$

$$F_{latency} = F_1 \cdot speed.$$

The individual penalty is given by a logistic function and depends on a scaling parameter, p:

$$penalty(k_u) = -1 + \frac{2}{1 + e^{-pk_u - 1}}$$

The scores attached to each subject are used to rank them. This ranking is evaluated by standard information retrieval metrics: $P@10$, $NDCG@10$ and $NDCG@100$. These are evaluated after 1, 100, 500 and 1000 writings have been processed.

2.3. Approaches and Training

Having no annotated training data at hand, data available on the web were exploited for both training and prediction. Two authorship attribution approaches were put forward for this task. In both cases, our approaches assess whether a test user belongs to a set of gambling testimonials using a similarity distance measure between their textual productions. A test user u is said to be at risk of pathological gambling if the minimal similarity distance δ^u_{min} computed for them is smaller than a threshold θ.

Since topic modeling have shown good potential in such authorship attribution task [8], it is selected to represent both test users' textual production and the testimonials. Given the performances of the Embedding Topic Model (ETM) [9] in [10], this model is selected for topic extraction. Our ETM model is trained on a corpus made from two datasets. The first part is made from the textual productions from the Subreddits *Problem Gambling*[1] and *Gambling Addiction Support*[2], ensuring the presence of gambling-related vocabulary in the corpus. The second part is made up of control subjects from the 2018 eRisk depression dataset, adding general topics to the corpus. Both gambling-related and control content were added in equal part to this novel training corpus in order to limit any discrepancies.

The ETM is trained following the methodology described in [10]. Using the trained model, the test user's textual productions and the testimonials are mapped to a vector of topic probabilities, which are then used in our two authorship attribution approaches to compute the similarity. Here, the similarity is given by computing the Hellinger distance between two vectors of topic probabilities.

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

[2] https://www.reddit.com/r/GamblingAddiction/

2.3.1. Testimonials

Our first approach consists in using testimonials found on the Web to assess the pathological gambling risk of the users from the test data. The testimonials offered by 199 compulsive gamblers were found on *Gambler's Help*[3]. These testimonials are considered to be our testimonials set $T = \{t_1, \ldots, t_{199}\}$.

Here, we aim to find the minimal similarity distance threshold θ_{min} to be considered at risk of pathological gambling by using the testimonials set T. Then, every testimonial $t \in T$ is compared to the others using a one-against-all cross validation technique, *i.e.* 1 vs. 198 testimonials, to compute its distance δ^t from every other testimonial. A testimonial t is represented by its vector of topic probabilities $\vec{t}$, which allows to compute the Hellinger distance between testimonials t_i and t_j, as

$$\delta^{t_i,t_j} = Hellinger(\vec{t_i}, \vec{t_j})$$

By doing so, it is possible to find the maximal similarity distance obtained for a testimonial compared to all the others, as

$$\delta^{t_i}_{max} = max(\{\delta^{t_i,t_j}, \ldots, \delta^{t_i,t_{n-1}}\})$$

Assuming that every testimonials has to be part of the testimonial set, the maximal similarity distance obtained across the evaluation is then the minimal threshold to be part of the set, as

$$\delta_{max} = max(\{\delta^{t_1}_{max}, \ldots, \delta^{t_n}_{max}\}) \qquad \theta_{min} = \delta_{max}$$

Thus, predicting if a test user pertains to the testimonial set is given by computing the similarity distance of its vector of topic probabilities against the vector of every testimonial. For a given test user, if the minimal similarity distance computed is lower than the threshold, then it is decided that the test user is part of the testimonial set.

$$\delta^{u}_{min} = min(\{\delta^{u,t_1}, \ldots, \delta^{u,t_n}\})$$

$$prediction(\delta^{u}_{min}, \theta_{min}) = \begin{cases} 1 & \text{if } \delta^{u}_{min} \leq \theta_{min} \\ 0 & otherwise \end{cases}$$

One potential issue with the use of these testimonials is that their language might differ from that used in Reddit fora. Nonetheless, topic models should smooth over the particulars by grouping word co-occurrences.

2.3.2. Questionnaire

Our second approach makes use of a self-evaluation questionnaire in addition to the set of testimonials. The self-evaluation questionnaire, which is often offered by resources for compulsive gamblers, was found on several websites, including *Gamblers Anonymous Montreal*[4]. This one is composed of 20 questions answerable by yes or no. An individual scoring 7 or more positive answers from this questionnaire is considered at risk of a pathological gambling problem.

[3] https://gamblershelp.com.au/

[4] http://gamontreal.ca/

Comparably to the testimonial approach, we aim to find the minimal similarity distance threshold θ_{min} to be considered at risk of pathological gambling. Here, this threshold is computed using the self-evaluation questionnaire and the testimonial set T. Given the questionnaire q and its vector of topic probabilities $\vec{q}$, a testimonial t is said close enough to the questionnaire to be considered at risk of pathological gambling if the Hellinger distance between $\vec{t}$ and $\vec{q}$ is less or equal to the threshold θ_{min}. Thus, the idea is to find the maximal similarity distance δ^{q}_{max} to define this threshold, as

$$\delta^{q}_{max} = max(\{\delta^{q,t_1}, \ldots, \delta^{q,t_n}\}) \qquad \theta_{min} = \delta^{q}_{max}$$

Then, predicting if a test user is at risk of pathological gambling can be made using its distance from the self-evaluation questionnaire, such as

$$prediction(\delta^{u,q}, \theta_{min}) = \begin{cases} 1 & \text{if } \delta^{u,q} \leq \theta_{min} \\ 0 & otherwise \end{cases}$$

2.4. Results

Table 1
Results obtained on the test set of Task 1 by our models and the best performing models on each metric. Runs are denoted APPROACH-stem and APPROACH-reg following the methodology adopted from [10]

System	Run	*precision*	*recall*	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	*speed*	$F_{latency}$
Questionnaire-stem	0	0.138	0.988	0.243	**0.048**	0.036	1	**1**	0.243
Questionnaire-reg	1	0.108	**1**	0.194	0.057	0.045	1	**1**	0.194
Testimonials-stem	2	0.071	**1**	0.132	0.067	0.064	1	**1**	0.132
Testimonials-reg	3	0.071	**1**	0.132	0.066	0.064	1	**1**	0.132
CeDRI	1	.070	**1**	.131	.066	.065	1	**1**	.131
UNSL	2	**.586**	.939	**.721**	.073	**.020**	11	.961	**.693**

Table 2
Ranking-based results ($P@10$; $NDCG@10$; $NDCG@10$) on the test set of Task 1 for our models and the best model

Team	Run	1 writing			100 writings			500 writings			1000 writings		
RELAI	0	.9	.92	.73	**1**	**1**	.93	**1**	**1**	.92	**1**	**1**	.91
	1	**1**	**1**	.72	**1**	**1**	.91	**1**	**1**	.91	**1**	**1**	.91
	2	.8	.81	.49	.5	.43	.32	.5	.55	.42	.5	.55	.41
	3	.8	.88	.61	.6	.68	.49	.7	.77	.55	.8	.85	.55
UNSL	2	**1**	**1**	**.85**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

The results are presented in Tables 1 and 2. Our best model was Run 0, outperforming our other approaches on both precision and F-measure. While showing a limited precision, it obtained the best $ERDE_5$ across every other system presented for this task.

3. Task 2: Early Detection of Signs of Self-Harm

The task was introduced in 2019, and teams did not have access to any training data, producing modest results [11]. The following year, Transformer-based approaches were the most prolific, achieving the best precision, F_1-score, $ERDE$s and latency-weighted F_1. XLM-RoBERTa models were trained on texts from the Pushshift Reddit Dataset [12], and predicted whether a user was at risk of self-harm or not by averaging on all their known posts. Each of their runs targeted a specific evaluation metric for the fine-tuning. As noted by [13], most runs had a near perfect recall and also a very low precision. Of those, NLP-UNED (runs 3 & 4) [14] seemed to gather the best overall performances. Their systems used a combination of textual features and sentiment analysis from the entire user's historic to predict whether they were at risk. The best results are presented in Table 3.

Table 3
Summary of the best results obtained on eRisk 2020 T1 (Self-harm).

System	Run	$precision$	$recall$	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$F_{latency}$
iLab [15]	0	0.833	0.577	0.682	0.252	0.111	10	0.965	**0.658**
	1	**0.913**	0.404	0.560	0.248	0.149	10	0.965	0.540
	2	0.544	0.654	0.594	**0.134**	0.118	2	0.996	0.592
	3	0.564	0.885	0.689	0.287	**0.071**	45	0.830	0.572
	4	0.828	0.692	**0.754**	0.255	0.255	100	0.632	0.476
NLP-UNED [14]	3, 4	0.246	**1**	0.395	0.213	0.185	**1**	**1**	0.395

3.1. Task and Data

The task objective and evaluation process are identical to that of Task 1, including the iterative evaluation of models and the metrics. The key difference, however, is that a training set was provided. The training set counted 145 positive subjects out of 763 (19.0%), while the test set counted 152 positive out of 1448 (10.5%).

3.2. Approaches

For this task, two approaches based on neural networks were tested. One is based on the Contextualizer encoder [16], while the other is based on RoBERTa embeddings [17].

3.2.1. Contextualizer

Following [18], two modes aggregating the different writings in a subject's history were used. The first, nested aggregation, uses one Contextualizer encoder to encode the writings separately into single vector representation and another Contextualizer encoder to aggregate writings together. The second mode, flat aggregation, performs both these steps at once by providing positional information to each word about its writing and within-writing position. For both of these approaches, the positional information about writings is not the chronological order but the time difference with respect to the most recent writing.

3.2.2. RoBERTa embeddings

A Transformer model was trained using RoBERTa [17]. This training was carried out on Reddit data by masked language modeling. This approach tokenizes writings into character n-grams based on their frequency in the source corpus. Once the Transformer was trained, the writings in the training set were transformed into token embeddings. These token embeddings constituting a writing, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are averaged together into a single document vector:

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

In order to combine these document representations into a single vector per subject, we posit that writings farther in the past should be given less importance than more recent ones. Given a set of m documents, $\{(\bar{\boldsymbol{x}}_j, t_j)\}_{j=1}^{m}$, where $t_j \in \mathbb{R}$ denotes the difference in hours from the j-th document to the most recent one, the document vectors are aggregated into a single vector:

$$\boldsymbol{u} = \sum_{j=1}^{m} \boldsymbol{\alpha}^{t_j} * \bar{\boldsymbol{x}}_j$$

Here, $\boldsymbol{\alpha}$ is a vector of learned parameters and the exponentiation and multiplication, $*$, are applied element-wise. This allows for each feature to decay at an independent rate. Thereafter, the predicted probability of having observed a positive instance is given by:

$$\hat{y} = \sigma(\boldsymbol{w}^\top \boldsymbol{u}),$$

where σ denotes the standard sigmoid function and $\boldsymbol{w}$ is a vector of learned parameters.

3.2.3. Training

All models are trained by gradient descent with a binary cross entropy minimization objective using the Adam algorithm [19]. To compensate for possible discrepancies between the proportions of labels between the training and test sets, a balanced validation set was built taking half of the positive subjects and a number of negative subjects to match. In addition, a stratified validation set was also tested. In training, subjects were inversely weighted in the loss function to account for the imbalance. For both approaches, different contiguous samples of writings from each subject are taken at each epoch. The size of such samples was chosen to allow models to make early decisions without requiring a long history of writings. In validation, however, the most recent documents for each subject are taken. Model selection was based on the area under the precision-recall curve, which is equivalent to the average precision. The selected models are presented in Table 4. Except for Run 5, all of the chosen models were validated on the balanced validation set.

3.3. Results

Classification and ranking-based results are presented in Tables 5 and 6, respectively. Label decisions were fairly quick and favored positive decisions, resulting in low $latency_{TP}$ (2 to 5). Run 2 notwithstanding, recall was high ($>$.85), resulting in low precision ($<$.25) and modest F_1

Table 4
Models selected for the test stage of Task 2. The number of writings indicates how many of the most recents writings per subject the model will consider.

Run	Model	Nb of writings
0	Flat Contextualizer	5
1	Nested Contextualizer	5
2	Roberta Embeddings	20
3	Roberta Embeddings	5
4	Roberta Embeddings (strat. validation)	5

Table 5
Classification results on the test set of Task 2 for our models and the best models per metric

Team	Run	$precision$	$recall$	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$F_{latency}$
RELAI	0	.138	.967	.242	.140	.073	5	.984	.238
	1	.114	.993	.205	.146	.086	5	.984	.202
	2	.488	.276	.353	.087	.082	2	.996	.352
	3	.207	.875	.335	.079	.056	2	.996	.334
	4	.119	.868	.209	.120	.089	2	.996	.206
Birmingham	0	**.757**	.349	.477	.085	.07	4	.988	.472
CeDRI	2	.116	**1.0**	.19	.096	.094	**1**	**1.0**	.19
UNSL	0	.336	.914	.491	.125	**.034**	11	.961	.472
	4	.532	.763	**.627**	.064	.038	3	.992	**.622**
UPV-Symanto	1	.276	.638	.385	**.059**	.056	**1**	**1.0**	.385

Table 6
Ranking-based results (P@10; $NDCG$@10; $NDCG$@10) on the test set of Task 2 for our models and the best models per metric

Team	Run	1 writing			100 writings			500 writings			1000 writings		
RELAI	0	.1	.06	.11	.4	.37	.46	.4	.32	.38	.5	.47	.41
	1	0	0	.12	.2	.12	.36	0	0	.27	.1	.06	.28
	2	.8	.71	.4	.4	.28	.4	**1**	**1**	.6	**1**	**1**	.57
	3	.7	.76	.43	0	0	.31	.9	.88	.59	.8	.75	.56
	4	.4	.44	.34	0	0	.21	.4	.34	.27	.5	.5	.31
UNSL	0	**1**	**1**	**.7**	.7	.74	**.82**	.8	.81	**.8**	.8	.81	**.8**
	4	**1**	**1**	.63	**.9**	.81	.76	.9	.81	.71	.8	.73	.69
UPV-Symanto	0	.8	.83	.53	**.9**	**.94**	.67	.9	.94	.67	0	0	0

($<$.34) for those runs. This is partly due to the smaller proportion of positive subjects in the test set. Run 2 achieved much higher precision than our other runs (.488) but at the price of low recall (.276) resulting in a comparable F_1 (.353). However, Run 2 seemed to outperform our other runs in ranking-based evaluation and achieving perfect P@10 and $NDCG$@10 with 500 and 1000 writings processed. Its $NDCG$@100 was also high, indicating an adjustment to the decision policy might benefit classification. Overall, as per the ranking-based metrics, the scores produced by our models seemed to improve from 100 to 1000 writings, with the exception of Run 1, which remained low throughout.

Table 7
Summary of the best results obtained on eRisk 2019 T3 (severity)

Run	AHR	ACR	ADODL	DCHR
CAMH_GPT_nearest_unsupervised [22]	23.81%	57.06%	**81.03%**	**45.00%**
UNSL [21]	**41.43%**	69.13%	78.02%	40.0%
	40.71%	**71.27%**	80.48%	35.00%

Table 8
Summary of the best results obtained on eRisk 2020 T2 (severity)

Run	AHR	ACR	ADODL	DCHR
BioInfo@UAVR [23]	**38.30%**	69.21%	76.01%	30.00%
iLab run2 [15]	37.07%	**69.41%**	81.70%	27.14%
prhlt_svm_features [24]	34.56%	67.44%	80.63%	**35.71%**
relai_lda_user [18]	36.39%	68.32%	**83.15%**	34.29%

4. Task 3: Measuring the severity of the signs of depression

As described by [13], the task consists in mapping a subject's writings to a well-known tool for the assessment of depression symptoms, the Beck Depression Inventory (BDI) [20]. In 2019, the two approaches that gathered the best performances leveraged the dependency between the severity of depression categories and the severity of the signs. The first aimed to predict the severity category and then deduce the severity of each sign of depression [21], achieving the most precise predicted answers to the BDI questionnaire. The second system leveraged textual similarity between the user's productions and the questions from the BDI questionnaire to fill it. By combining those answers, the best results regarding the prediction of depression severity were obtained [22]. The results are presented in Table 7. A description of each evaluation metric is provided at Section 4.2.

For the second iteration of this task in eRisk 2020, the best performances remained similar to those observed in the previous year. The approaches achieving the best results were based on psycholinguistic features [23], pre-trained Transformers [15], LDA-based authorship attribution [18], or combining a support vector machine with a radial basis kernel [24]. The 2020 best results are presented in Table 8.

4.1. Task and Data

As with Tasks 1 and 2 the dataset comprises a history of writings per subject. However, instead of a binary label, each subject is associated with a set of 21 labels corresponding to the answers they gave to each item of the BDI. Furthermore, evaluation did not include a temporal aspect: the entire history of writings for the test subjects was made available at once. As shown in Fig. 1 the BDI scores are overall higher in the test set, with the median and median absolute deviation for the training and test set being (20.0, 9.5) and (27.0, 10.0) respectively.

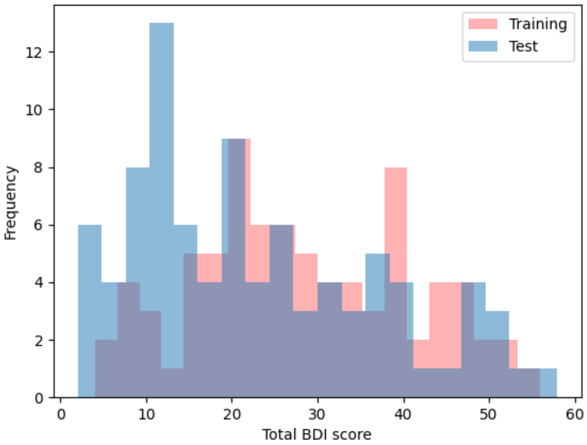

Figure 1: Histogram of total BDI scores in the training and test sets

4.2. Evaluation

In order to evaluate BDI predictions against the true BDI answers associated with a set of subjects, [13] propose four metrics. The Average Hit Rate (AHR) is the rate of exactly correct predictions averaged across the 21 items of the BDI and across subjects. In contrast, the Average Closeness Rate (ACR) measures the proximity in value ([0,3]) between the predicted and true answer when compared to the maximum possible difference (3). Similarly, the Average Difference in Overall Depression Levels (ADODL) compares the total score of the predicted BDI to the true total score, once again normalized by the maximum (63). Finally, the Depression Category Hit Rate (DCHR) is the accuracy in the depression categorization resulting from the predicted BDIs of subjects.

4.3. Approaches

Following [8], we opt for predicting the BDI items based on the similarities of the writings of the test subjects to those of the training subjects. These similarities are computed on learned representations of the textual production of subjects. These representations were based on topic modeling or neural encoders trained on authorship decision. In addition to the categorical prediction of BDI items proposed by [8], a regression approach (Reg.) based on the values of the answers was also tested, with the values being multiplied by the relevant similarity score. There is a high variance in answers even among subjects in the same depression category. To address this in the regression approach, each training subject's answer to each question is smoothed to the average answer in their depression category by a hyperparameter, $\beta \in [0, 1]$,

$$a_j \leftarrow \beta * a_j + (1 - \beta) * \bar{a}_j$$

Here, a_j denotes the answer selected by the j-th training subject to a given item in the BDI and $\bar{a}_j$, the average answer selected by the subjects in the depression category that said subject belongs too.

Variance aside, this approach is still potentially highly sensitive to the particular distribution of BDI scores in the training set. To address this, a nearest-neighbors approach was tested wherein a set number of neighbors was to be drawn from each of the four depression categories. This approach is denoted k'NN, and can be applied in both the regression and categorical settings.

4.3.1. Topic Modeling

Topic modeling consists in inferring probability distributions over a vocabulary of words, such that the documents, the subjects' histories in our case, constitute a mixture of such distributions. As a baseline, we selected the well-known Latent Dirichlet Allocation (LDA) algorithm. Further, another topic model, operating on word embeddings rather than symbolic word representations like LDA, ETM [9], was also tested. Models were trained on a depression-detection dataset also issuing from Reddit [25]. This training was carried out considering the entire history of writings from each subject as a single document. For the LDA approach, two tokenization schemes were tested: character trigrams and word stems. In contrast, only stemming was tested for the ETM model for interpretability purposes.

4.3.2. Authorship decision

Deep Averaging Networks (DANs) were trained to discern whether two sets of writings were authored by the same person. This can induce a representation relevant to depression symptoms [8]. As with topics models, these models were trained on the eRisk 2018 Depression dataset, using alternately character trigrams and word stems. The training procedure consists in sampling non-overlapping sets of writings from subjects and pairing them together. Pairs of samples issuing from the same subject constitute positive examples and pairs issuing from different subjects, negative ones.

4.3.3. Model selection

As previously mentioned, this approach is potentially highly sensitive to the distribution of BDI scores in the training set. In order to mitigate this, the validation set selected contained 24 subjects equally divided among the four depression categories defined by the BDI. The hyper-parameter values tested were borrowed from [8].

Models were selected based on the performance on all four metrics. Indeed, selecting the top performers for each metric separately might exclude models performing well overall. However, in order to combine all four metrics into a single quantity by which to select models requires consideration. Although the metrics are valued in the unit interval, they have different scales in practice. Therefore, combining the performance for each metric for all models and hyper-parameter values, the z-score for each one was computed. Then, the average z-score across all metrics was used to select the models. The selected models are shown in Table 9. As in [8], k denotes the number of neighbors considered, while δ is the consensus parameter of DMkNN. D and t denote the size of the writing history partition and the number of parcel pairs considered.

Table 9
Models selected for prediction on the test set of Task 3

Run	Encoder	Algorithm	Tokenization	k	δ	D	t
0	DAN	DMkNN	trigram	30	10	1	1
1	DAN	DMkNN	trigram	9	7	10	1
2	ETM	Reg. k'NN	stemming	5	-	1	1
3	DAN	k'NN	trigram	5	-	1	1
4	LDA	Reg. k'NN	trigram	3	-	10	10

Table 10
Results (%) on the test set of Task 3 for our models and the best models per metric

Team	Run	AHR	ACR	ADODL	DCHR
RELAI	0	34.64	67.58	78.59	23.75
RELAI	1	30.18	65.26	78.91	25.00
RELAI	2	**38.78**	72.56	80.27	35.71
RELAI	3	34.82	66.07	72.38	11.25
RELAI	4	28.33	63.19	68.00	10.00
DUTH ATHENA	5	**35.36**	67.18	73.97	15.00
UPB	5	34.17	**73.17**	82.42	32.50
CYUT	2	32.62	69.46	**83.59**	**41.25**

4.4. Results

The results are shown in Table 10. Our best model was Run 2, outperforming the rest of our models on each metric. Overall, the total BDI scores predicted were low, with Run 0 having the highest median of 18 and Run 3 having the lowest of 8. Predictions were quite tight within each run, with Run 1 having the highest median absolute deviation of 6. Furthermore predictions were consistent among runs, with Run 1 and 4 agreeing the least, on only 44% of answers globally. Interestingly, although Run 0 agreed the most with Run 2 (64%), it achieved much weaker results, especially in terms of DCHR.

Overall, the approach remains sensitive to the particulars of the training set where neighbors are sourced. This is perhaps due to text alone not eliciting, by unsupervised learning alone, similarity that pertains to depression symptoms. Future work could include integrating manual annotation or prior knowledge in the training of the similarity models, authorship and topic alike. Moreover, in order for the overall approach to be effective in the 21-way prediction at hand, similarity could be handled separately by component or groups of components of the subject representation.

5. Conclusion

RELAI participated in all three eRisk 2021 shared tasks. Task 1, *Early Detection of the Signs of Pathological Gambling*, proved an interesting challenge in the lack of training data. Nonetheless, the use of testimonials and self-assessment questionnaires constitutes a promising avenue in such a context. The *Early Detection of the Signs of Self-Harm*, Task 2, was a more conventional one. The favoring of early decisions resulted in high recall but poor precision overall. Nonetheless, some of the proposed approaches produced good ranking-based results. Finally, Task 3, *Measuring the Severity of the Signs of Depression*, remains thoroughly difficult. However, in predicting BDI scores based on similarity, restricting the number of neighbors per depression category proved an interesting option to address uncertain distributions of BDI scores.

The source code of the proposed systems is licensed under the GNU GPLv3. The datasets are provided on demand by the eRisk organizers.

Acknowledgments

This research was enabled by support provided by Calcul Québec and Compute Canada. MJM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada [NSERC Grant number 06487-2017] and the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFE-2018-00484].

References

[1] J. Parapar, M.-R. Patricia, D. E. Losada, F. Crestani, Overview of eRisk 2021: Early Risk Prediction on the Internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2021.

[2] M. W. Abbott, The changing epidemiology of gambling disorder and gambling-related harm: Public health implications, Public health (2020).

[3] S. Achab, A. Chatton, R. Khan, G. Thorens, L. Penzenstadler, D. Zullino, Y. Khazaal, Early detection of pathological gambling: betting on GPs' beliefs and attitudes, BioMed research international 2014 (2014).

[4] J. Haefeli, S. Lischer, J. Schwarz, Early detection items and responsible gambling features for online gambling, International Gambling Studies 11 (2011) 273–288.

[5] J. Braverman, D. Laplante, S. Nelson, H. Shaffer, Using cross-game behavioral markers for early identification of high-risk internet gamblers, Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors 27 (2013) 868–77. doi:10.1037/a0032818.

[6] J. Haefeli, S. Lischer, J. Haeusler, Communications-based early detection of gambling-related problems in online gambling, International Gambling Studies 15 (2015) 23–38. URL: https://doi.org/10.1080/14459795.2014.980297. doi:10.1080/14459795.2014.980297. arXiv:https://doi.org/10.1080/14459795.2014.980297.

[7] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 495–503.

[8] D. Maupomé, M. D. Armstrong, F. Rancourt, M.-J. Meurs, Leveraging textual similarity to predict beck depression inventory answers, Proceedings of the Canadian Conference on Artificial Intelligence (2021). URL: https://caiac.pubpub.org/pub/pkzbt8x2. doi:10.21428/594757db.5c753c3d, https://caiac.pubpub.org/pub/pkzbt8x2.

[9] A. B. Dieng, F. J. Ruiz, D. M. Blei, Topic Modeling in Embedding Spaces, Transactions of the Association for Computational Linguistics (2020).

[10] M. D. Armstrong, D. Maupomé, M.-J. Meurs, Topic modeling in embedding spaces for depression assessment, Proceedings of the Canadian Conference on Artificial Intelligence (2021). URL: https://caiac.pubpub.org/pub/b6tk9kak. doi:10.21428/594757db.9e67a9f0, https://caiac.pubpub.org/pub/b6tk9kak.

[11] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019: Early Risk Prediction on the Internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.

[12] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The Pushshift Reddit Dataset, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 830–839.

[13] D. E. Losada, F. Crestani, J. Parapar, eRisk 2020: Self-harm and depression challenges, in: European Conference on Information Retrieval, Springer, 2020, pp. 557–563.

[14] E. C. Ageitos, J. Martínez-Romo, L. Araujo, NLP-UNED at eRisk 2020: Self-harm Early Risk Detection with Sentiment Analysis and Linguistic Features, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2696, 2020.

[15] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using BERT-based transformers: iLab at CLEF eRisk 2020, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2696, 2020.

[16] D. Maupomé, M.-J. Meurs, An Iterative Contextualization Algorithm with Second-Order Attention, arXiv preprint arXiv:2103.02190 (2021).

[17] Y. Liu, M. O. andhttps://latex.ikb.info.uqam.ca/project/600f60b981c0730096ee0382 Naman Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint 1907.11692 abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[18] D. Maupomé, M. D. Armstrong, R. Belbahar, J. Alezot, R. Balassiano, M. Queudot, S. Mosser, M.-J. Meurs, Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2696, 2020.

[19] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, CoRR abs/1412.6980 (2014). URL: http://arxiv.org/abs/1412.6980. arXiv:1412.6980.

[20] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An Inventory for Measuring Depression, Archives of General Psychiatry 4 (1961) 561–571. doi:10.1001/archpsyc.1961.01710120031004.

[21] S. G. Burdisso, M. Errecalde, M. Montes y Gómez, UNSL at eRisk 2019: a unified approach for anorexia, self-harm and depression detection in social media, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2380, 2019.

[22] P. Abed-Esfahani, D. Howard, M. Maslej, S. Patel, V. Mann, S. Goegan, L. French, Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2380, 2019.

[23] A. Trifan, L. Salgado, Pedro aand Oliveira, BioInfo@ UAVR at eRisk 2020: on the use of

psycholinguistics features and machine learning for the classification and quantification of mental diseases, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2696, 2020.

[24] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, volume 2696, 2020.

[25] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 343–361.