Automatically Finding Evidence and Predicting Answers in Mental Health Self-Report Questionnaires

Diego Maupomé[1], Yves Ferstler[1], Sébastien Mosser[2] and Marie - Jean Meurs[1]

[1]*Université du Québec à Montréal, QC, Canada*
[2]*McMaster University, ON, Canada*

Abstract

This paper describes the participation of the RELAI team in the eRisk 2024 shared tasks related to the search for symptoms of depression (T1) and the measure of severity of the signs of eating disorders (T3). Both tasks centered on self-report questionnaires and social media textual content. In T1, sentences relevant to each item in a standard depression were mined from a large set of sentences collected from social media. Our approach relied on ground-truth relevance judgments to train multi-label classifiers in deciding whether a sentence is relevant to each item. The goal of T3 was to automatically fill out a standard eating disorder questionnaire based on histories of writings from social media. Given the small set of annotated data and large output space, our approach proceeded by making global, aggregate predictions first and use those predictions to make precise per-item predictions.

Keywords

Mental Health, Mental Disorders, Depression, Eating Disorders, Natural Language Processing,

1. Introduction

According to the WHO fact sheets, around 5% of adults worldwide suffer from depression in 2023[1], and 14 million people experienced eating disorders in 2019[2]. Since the number of social media users worldwide reached more than 5 millions in 2024, *i.e.* almost 63% of the world's population, it is still critical to contribute to early detection of mental disorders by developing scalable tools based on natural language processing and machine learning, which can automatically analyze online textual content shared by social media users.

The 2024 edition of the eRisk evaluation campaign proposed three tasks respectively focusing on the search of depression symptoms (T1), the early detection of signs of anorexia (T2) and the estimation of severity of signs of eating disorders (T3) [1, 2]. T1 was the continuation of the similar task created in 2023, though teams were provided this year with a ground truth set of tagged sentences created for last year task. T2 was a continuation of eRisk 2018 and 2019 tasks. T3 was a continuation of 2022 and 2023 eRisk tasks. This paper describes the participation of the RELAI team to tasks 1 and 3, using lightweight approaches to minimize computation costs and infrastructure needs, as in most of our previous work [3, 4, 5].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09-12, 2024, Grenoble, France
maupome.diego@courrier.uqam.ca (D. Maupomé); ferstler.yves@courrier.uqam.ca (Y. Ferstler); mossers@mcmaster.ca (S. Mosser); meurs.marie-jean@uqam.ca (M. -. J. Meurs)
0000-0003-2527-2515 (D. Maupomé); 0009-0000-2982-1281 (Y. Ferstler); 0000-0001-9769-216X (S. Mosser); 0000-0001-8196-2153 (M. -. J. Meurs)

[1]WHO Fact Sheet - Depressive disorder (depression), 31 March 2023, accessed on 31 May 2024
[2]WHO Fact Sheet - Mental disorders, 8 June 2022, accessed on 31 May 2024

In T1, the goal of the system is to retrieve relevant sentences, and rank them according to their relevance to signs and symptoms of depression described in the Beck Depression Inventory, 2nd edition (BDI-II) [6]. Our approaches make use of ground-truth relevance judgments to train multi-label classifiers to predict the relevance of sentences to each BDI item.

In T3, the system must fill questions 1-12 and 19-28 of the Eating Disorder Examination Questionnaire (EDE-Q) [7], in order to automatically estimate the severity of signs of eating disorders. Given this complexity of predicting 22 variables and limited availability of training data, our approach makes use of global and subscale scores: beginning with predicting broader aggregate scores and relying on this output to make more precise, item-level predictions.

2. Task 1: Search for symptoms of depression

2.1. Task and Data

The task consists in retrieving and ranking sentences based on their relevance to signs and symptoms of depression, as identified by the Beck Depression Inventory (BDI) [6]. This self-report questionnaire inventories affective, cognitive, somatic and vegetative symptoms of depression in 21 items to provide a measure of the overall severity of depression.However, the task is not concerned with the severity of these symptoms, consisting instead in finding sentences relevant to each of these items from a common pool of several million sentences from Internet fora. For each of the 21 items, a relevance ranking of up to 1000 sentences is expected. Sentences are allowed to appear in several item rankings. Relevance is defined as the presence of both pertinence to the given symptom and "explicit information about the individual's state in relation to it" [8].

Training data consists of 3.8M sentences, among which 16.1k have been manually assessed for relevance. Relevance judgments are provided based on a majority as well as the unanimity of the three annotators. Statistics on these relevance judgments are presented in Table 1. Sentences assessed for relevance were pooled from the rankings provided by task participants in the previous iteration of the task (2023). The test set contains 15.5M sentences.

As shown in Table 1 around 16.1k unique sentences were assessed for relevance over 21.6k attributions across items. For both majority and consensus relevance, a small proportion (resp. 11%, 22%) of these attributions were deemed relevant overall. The number of relevant sentences varies by an order of magnitude across items (Figure 1a). Furthermore, relevance to several items decreases exponentially in the number of items (Figure 1b).

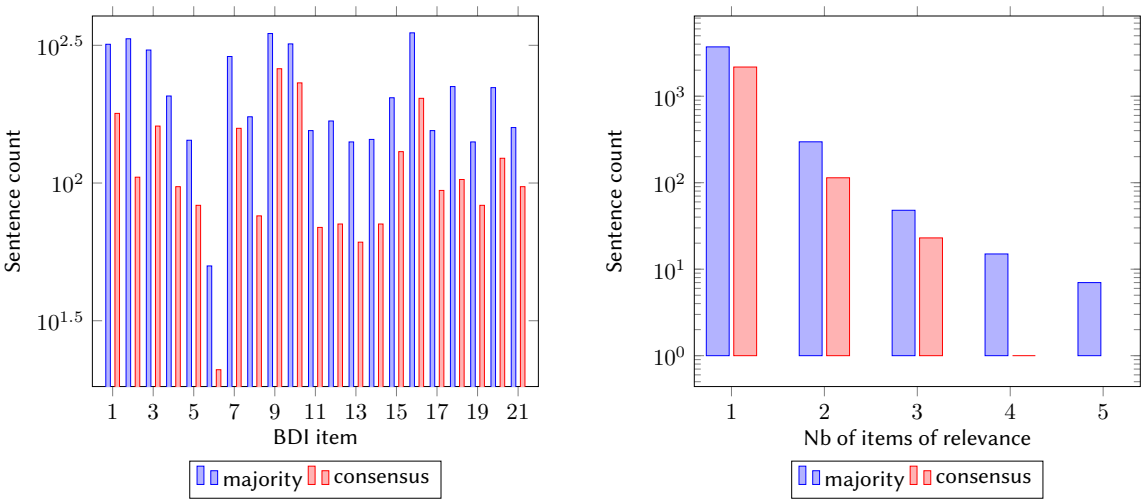
Task performance was evaluated using standard information retrieval performance measures, namely average precision (AP), top-ten precision (P@10), r-precision (R-P) and non-discounted cumulative gain (NDCG) [2].

2.2. Approaches and training

Participants in the previous iteration of this task did not have access to ground-truth relevance judgments. Past approaches are therefore completely unsupervised. These approaches are based on the use of textual similarity between target sentences and reference texts. Choices for the latter vary between the text of the questionnaire itself, paraphrases and expansions thereof and expert features. By and large, the measure of similarity is based on pretrained transformer approaches, with coarser-grained filtering based on feature words applied on occasion.

Table 1
Summary counts on relevance judgments for the training set of Task 1: Search for symptoms of depression

	consensus	majority
annotations	21 580	21 580
unique annotated	16 148	16 148
relevant	2 476	4 552
unique relevant	2 313	4 086



(a) Number of sentences labeled as relevant per BDI item (b) Frequency of multiple relevances per sentence

Figure 1: Relevant sentence counts for different items as per majority and consensus rules.

However, such approaches rely on the design of relevant reference texts for items or the assumption that textual similarity to the questionnaire entails positive relevance as defined by the task. Given these difficulties capturing relevance through unsupervised means and the availability of relevance judgments, our approach to the task was to induce relevance in a supervised manner, namely through classification. The task was framed as a multilabel classification task with each item acting as a label. The scores produced by the resulting probabilistic models were then used as relevance scores for ranking purposes.

These multilabel classification models were feed-forward neural networks operating on transformer embeddings of the input sentences. That is, sentences were encoded using pretrained neural sentence encoders. From this representation, feed-forward networks were trained to predict the relevance of sentences to all items of the BDI in a joint manner.

An initial concern in training these networks was the effect on model conditioning of the large number of overall negative sentences (not relevant to any item) in the dataset. To measure these effects, validation experiments controlled the proportion of overall negative sentences in training. The effects observed were contrary to initial intuitions, with a larger number of negative examples yielding better results across all metrics. Indeed, as shown in Table 2, results appear to improve overall as the number of negative sentences in training increases, likely due to the increase in total number of examples. Only in using all negative examples as per majority annotation does performance decrease with respect to

Table 2
Validation results for different controlled ratios of overall negative to positive sentences for a feed-forward network with one hidden layer operating on sentence embeddings from MiniLM [9]. "Corpus" indicates the full set of relevance judgments was used.

negative-to-positive ratio	rel.	AP	P@10	R-P	NDCG
0.5	majority	0.216	0.367	0.250	0.636
1	majority	0.240	0.457	0.255	0.671
2	majority	0.280	0.538	0.317	0.701
corpus (4.7)	majority	0.267	0.457	0.295	0.688
0.5	consensus	0.0	0.0	0.0	0.0
1	consensus	0.201	0.300	0.251	0.548
2	consensus	0.224	0.357	0.243	0.583
corpus (7.7)	consensus	0.248	0.381	0.262	0.607

Table 3
Test results for Task 1: Search for symptoms of depression (majority voting). Our models are compared to the best models for each performance measure, in bold.

model	AP	P@10	R-P	NDCG
paraphrase-MiniLM-L12-v2	0.267	0.346	0.738	0.525
paraphrase-MiniLM-L6-v2	0.236	0.325	0.590	0.503
all-MiniLM-L6-v2-simcse	0.226	0.322	0.595	0.495
tfidf sgd	0.163	0.240	0.552	0.394
NUS-IDS Config 5	**0.375**	**0.434**	0.924	**0.631**
MeVer-REBECCA TransformerEmbeddings CosineSimilarity gpt	0.301	0.340	0.981	0.506
APB-UC3M sentsim-all-MiniLM-L6-v2	0.354	0.391	**0.986**	0.591

the highest controlled ratio. In light of this, the number of negative sentences was not controlled during the training of models for submission.

Model selection was done by performing grid search on the choice of sentence encoder and depth of the feed-forward classifier, using a 40% validation set. The sentence encoders evaluated were several variants and sizes of Mini-LM, namely the variants trained on paraphrasing as well as the default, all-task variants [9]. The number of hidden layers test varied from one to three. Models were ranked by their average z-score across metrics with respect to other models [10]. As a baseline classification-based approach, we used term frequency representations with a logistic regression model.

2.3. Results

Results on the test set are shown in Tables 3 and 4. Our transformer-based approaches outperformed the baselines overall and performed consistently with validation results. However, they were outperformed by top participating runs across all metrics, especially in r-precision. Whether these differences are due to reduced capacity (the only parameters are those of the classifier) or the premise of the approach is difficult to establish without further experimentation. Within our transformer-based approaches, the largest model (paraphrase MiniLM with 12 layers) obtained the best results overall. No marked difference was observed between the two smaller models that differ only in their pretraining task.

Table 4
Test results for Task 1: Search for symptoms of depression (consensus). Our models are compared to the best models for each performance measure, in bold.

model	AP	P@10	R-P	NDCG
paraphrase-MiniLM-L12-v2	0.248	0.329	0.576	0.537
paraphrase-MiniLM-L6-v2	0.207	0.287	0.410	0.509
all-MiniLM-L6-v2-simcse	0.194	0.275	0.433	0.499
tfidf sgd	0.138	0.207	0.376	0.383
NUS-IDS Config 5	**0.392**	**0.436**	0.795	**0.692**
MeVer-REBECCA TransformerEmbeddings CosineSimilarity gpt	0.305	0.357	**0.833**	0.551
APB-UC3M sentsim-all-MiniLM-L6-v2	0.345	0.407	0.829	0.630

Further work could involve layer-wise finetuning to increase capacity.

3. Task 3: Measuring the severity of the signs of eating disorders

3.1. Task and Data

The third task of eRisk 2024 consists in measuring the severity of the signs of eating disorders. These signs are selected questions coming from the Eating Disorder Examination Questionnaire (EDE-Q) [7], a list of 22 questions of a scale from 0 to 6 included. These questions are grouped into four subscales concerning different aspects of eating disorders : restraint and eating, shape and weight concern. Scores for each question are averaged to obtain a score for each subscale, which are in turn used to compute a global score. Thus, the goal of task 3 is to predict the severity of eating disorder symptoms by forecasting the responses to the EDE-Q questionnaire from users.

The dataset matches partial histories of writings of Reddit users to their set of answers to the EDE-Q. The data for each user contains an id and a list of posts they made. The number of posts per user can vary. Each post contains a title and a body holding the content from the post. Notice that a post without title specified will be considered as a response. The combination of the data from 2022 and 2023 was used as training data, for a total of 74 examples. The test set contained 18 examples.

Models were evaluated using eight different metrics based on proximity to the true answers given by a subject to the EDE-Q. The first three evaluation metrics are based on the scores of all the questions from the psychometric scale independently. The five following metrics are concerned with the error in total scores resulting from the answers provided by the model, with possible focus on particular subscales. These metrics are presented below:

Mean Zero-One Error (MZOE): Calculate the error rate between the questionnaire filled by the user and predicted by the system.

$$MZOE(f, Q) = \frac{|\{q_i \in Q : R(q_i) \neq f(q_i)\}|}{|Q|}$$

Where f represent the classifier, Q the set of question from the questionnaire and the function $R(q_i)$ the real user answers from the question i-th. If a predicted answer form the system doesn't match with the real answer, the value will be equal to 1, and 0 otherwise. **Mean Absolute Error** (MAE): Calculate

the deviation between the questionnaire filled by the user and predicted by the system.

$$MAE(f, Q) = \frac{\sum_{q_i \in Q} |R(q_i) - f(q_i)|}{|Q|}$$

Instead of a binary evaluation, the metric calculate the discrete difference between the system and the real user answers. Knowing that the response can go from 0 included to 6 included.

Macroaveraged Mean Absolute Error ($\text{MAE}_{\text{macro}}$): Calculate the deviation between the questionnaire filled by the user and predicted by the system.

$$MAE_{\text{macro}}(f, Q) = \frac{1}{7} \sum_{j=0}^{6} \frac{\sum_{q_i \in Q_j} |R(q_i) - f(q_i)|}{|Q_j|}$$

Similarly to the MAE metric, $\text{MAE}_{\text{macro}}$ calculates the deviation by grouping the questions with the same answer together.

Restraint Subscale (RS): calculates the mean response of the question from the restraint subscale only. The mean is calculated between the questionnaire filled by the user and predicted by the system.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} \left(R_{RS}(u_i) - f_{RS}(u_i)\right)^2}{|U|}}$$

where U represent the set of question relative to the specific subscale from the questionnaire. $R_{RS}(u_i)$ represent the estimated restraint subscale RS score and $f_{RS}(u_i)$ the real one.

Eating Concern Subscale (ECS): calculates the mean response of the question from the eating concern subscale only.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} \left(R_{ECS}(u_i) - f_{ECS}(u_i)\right)^2}{|U|}}$$

where $R_{ECS}(u_i)$ represent the estimated eating concern subscale ECS score and $f_{ECS}(u_i)$ the real one.

Shape Concern Subscale (SCS): calculates the mean response of the question from the shape concern subscale only.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} \left(R_{SCS}(u_i) - f_{SCS}(u_i)\right)^2}{|U|}}$$

where $R_{SCS}(u_i)$ represent the estimated shape concern subscale SCS score and $f_{SCS}(u_i)$ the real one.

Weight Concern Subscale (WCS): calculates the mean response of the question from the weight concern subscale only.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} \left(R_{WCS}(u_i) - f_{WCS}(u_i)\right)^2}{|U|}}$$

where $R_{WCS}(u_i)$ represent the estimated weight concern subscale WCS score and $f_{WCS}(u_i)$ the real one.

Global ED (GED): Finally the global score is calculated by making the mean of all the subscale.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} \left(R_{GED}(u_i) - f_{GED}(u_i)\right)^2}{|U|}}$$

3.2. Approaches and Training

3.2.1. Past work

Task 3 is in its third iteration, having taken place in 2022 and 2023. However, only participants to the 2023 iteration had accessed to labeled data, with 2022 participants having to propose completely unsupervised approaches. Thus, we examine the approaches proposed in 2023 [8]. It should be noted that the approach proposed by GMU-FAST is omitted from this treatment, as a detailed methodology is not available.

BFH-AMI used first the GPT-2 [11] model to embed the questions from the psychometric scale EDE-Q and the content of the users post, then used this embedding to do the prediction with a logistic regression model.

UMU used manual prepossessing techniques removing contractions, mentions, hashtags, URLs, and AMP expressions. Additionally, they used emoji extraction on the data. Finally they fine-tuned the model multi-qa-mpnet-base-dot-v1 which is based on MPNet to achieve their result.

Finally RiskBusters used a topic-driven approach in order to solve the task. By using BERTopic [12] framework to extract a list of topics from the user's posts and supported by a domain-adapted model MentalBERT [13], RiskBusters achieved the best performance, GMU-FAST notwithstanding.

3.2.2. Challenges

There are two main challenges that persist in Task 3 of eRisk. The first one is the complexity of the expected output. Indeed, the objective of this task is to predict with an automated system the answers form the questionnaire EDE-Q, i.e. predict 22 questions with answers ranging from 0 to 6. Making these predictions jointly complexifies the model and sparsifies the available data. In contrast, treating the prediction of each question as an independent task ignores the interactions between them.

Another challenge in this task is the small amount of data: 74 labelled examples. This limitation increases the difficulty to have a robust system. Moreover, the sum of posts for some users can be low and not relevant for the process. Thus, throughout this project, it will be important to address these two main challenges in addition to any others that may arise.

3.2.3. Design

In order to address these challenges, our approach leverages the structure of the EDE-Q. Indeed, our approach avoids directly predicting all items jointly, by predicting more global scores over several steps:

1. **Predict global score:** the total score resulting from the entire set of answers is predicted from the input writing histories.
2. **Predict subscale scores:** the aggregate scores for each subscale is predicted jointly from the input as well as the total predicted global score
3. **Predict answers to questions :** Finally, the last stage aims to predict the answer for each question based on the input and the aggregate predictions of the previous steps.

By proceeding in this manner, we created a dependence between each question. The final score of a question will not only depend on the user's posts but also on stage 1, which estimates whether the user shows signs of an eating disorder, and stage 2, which estimates the concern score for the question.

Further, the global score of the EDE-Q, which ranges between 0 and 6, can be compared against a threshold of 3, which indicates pathological severity. Extending this threshold to subscales and

questions, each stage of prediction can be transformed into a binary classification task. A priori, this transformation can be applied independently to each stage. To produce answers from classification models, equal-width probability strata are used.

3.2.4. Detailed Model

Given the variable size of writing histories, an appropriate representation needs to be built to feed into the model. We decided to use BERTopic [12] to extract topics from each user and use these topics as features for prediction.

Once the list of topics is found for each user, we selected the 16 topics with the most correlation with signs of eating disorder. In order to know which topics are the most important, a random forest was trained, and the most salient features were selected.

Prediction models were feed-forward neural networks taking as input topic probabilities as well as the applicable broaded score prediction. Several combinations of stage approaches (classification or regression) were tested in validation. Further, sharing layers between stages was also evaluated.

3.3. Results

Five runs have been submitted for the evaluation. For the two first runs, all the stages have been framed as classification. The second run uses a shared encoder between these tasks, whereas the first one and all the other runs have separate encoders. The third run uses classification for the first two stages and regression for its last stage, and the run 4 only has been trained as a regression matter.

Results on the test are shown in Table 5. The scores obtained from the 5 runs indicate that the model trained with classification generally achieved the lowest errors globally. Notably, run 3 also produced low results, particularly in the MZOE and ECS metrics.

However, run 1 shows the best result in the WCS metric. Additionally, when compared to the baseline results, the model scores are consistently higher.

Table 5
Test results for Task 3: Measuring the severity of the signs of eating disorders. Our models are compared to the best models for each performance measure, in bold.

team	run ID	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
baseline	all 0s	3.790	0.813	4.254	4.472	3.869	4.479	4.363	3.361
baseline	all 6s	1.937	0.551	3.018	3.076	3.352	2.868	3.029	2.472
baseline	average	1.965	0.594	3.137	2.875	3.361	2.102	2.229	2.306
RELAI	0	2.331	0.914	2.243	2.394	2.222	2.324	2.340	**1.812**
RELAI	1	2.346	0.917	2.237	2.507	2.199	2.216	2.328	1.836
RELAI	2	2.758	0.934	2.885	2.883	2.767	3.126	3.061	2.171
RELAI	3	2.356	0.775	2.233	2.700	2.928	3.266	2.106	2.310
RELAI	4	2.851	0.884	2.979	3.159	2.784	3.150	3.068	2.336
SCaLAR-NITK	0	1.912	0.591	**1.643**	2.495	2.713	1.568	**1.536**	2.098
SCaLAR-NITK	1	1.980	0.664	1.972	2.570	2.562	**1.553**	1.960	2.066
SCaLAR-NITK	2	1.879	**0.568**	1.942	2.158	2.477	2.222	2.245	2.364
SCaLAR-NITK	3	1.932	0.586	1.868	**2.117**	2.430	2.046	2.242	2.407
SCaLAR-NITK	4	**1.874**	0.672	1.820	2.292	**2.140**	1.557	1.880	2.061

Overall, the performance of the model requires significant improvement in the independent questions areas. Despite this, the model shows encouraging results with the WCS metric, indicating competitive performance in this specific area.

4. Conclusion

Our participation in eRisk 2024 was centered around self-report questionnaires. In T1, the goal was to mine sentences relevant to each item in the Beck Depression Inventory from a pool of sentences from social media. Our approach made use of ground-truth relevance judgments to train neural classifiers to predict the relevance of sentences to items in a multi-label setup. This approach obtained consistent results in validation and testing that were ultimately weaker than other approaches proposed in the shared task. Future work could involve combining both unsupervised, similarity-based methods of ranking with supervised means.

In contrast, T3 requires to fill out a standard eating disorder severity questionnaire, EDE-Q, based on partial histories of writings of social media users. Given the complexity of prediction and limited access to annotated data, our approach used staggered models with increasingly specific predictions, starting with global score prediction and ending with specific answer prediction. This approach yielded modest results, outperforming shared task baselines but remaining shy of leading approaches. However, we sustain the core principle of our approach as a sound means to leverage the structure built into the design of the questionnaire. There is vast room for improvement within this approach, from exploring different writing history representations to improving model calibration when applying a classification transformation.

Acknowledgments

The authors thank Sublime Tshimpangila, Jessica Dawson and Matthew Curtis for the fruitful discussions. Ayed Naber, Benjamin Chinnery, Michael Breau, Mohamed Aly, Rylan Sykes, Shrill Patel, Yaruo Tian, Yili Liu are also thanked for their interest in the RELAI team's work. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [MJ Meurs, NSERC Grant number 06487-2017].

References

[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024, Springer International, Grenoble, France, 2024.

[2] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early risk prediction on the Internet (Extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, CEUR Workshop Proceedings, 2024.

[3] D. Maupomé, T. Soulas, F. Rancourt, G. Cantin-Savoie, G. Winterstein, S. Mosser, M.-J. Meurs, Lightweight methods for early risk detection, Working Notes of CLEF (2023) 18–21.

[4] S. H. H. Saravani, L. Normand, D. Maupomé, F. Rancourt, T. Soulas, S. Besharati, A. Normand, S. Mosser, M.-J. Meurs, Measuring the severity of the signs of eating disorders using similarity-based models., in: CLEF (Working Notes), 2022, pp. 936–946.
[5] M. D. Armstrong, D. Maupomé, M.-J. Meurs, Topic modeling in embedding spaces for depression assessment, in: Proceedings of the Canadian Conference on Artificial Intelligence, 2021. doi:10.21428/594757db.9e67a9f0.
[6] A. T. Beck, R. A. Steer, G. K. Brown, Beck Depression Inventory (BDI-II), Psychological assessment 10 (1996).
[7] J. M. Mond, P. J. Hay, B. Rodgers, C. Owen, P. J. Beumont, Validity of the eating disorder examination questionnaire (ede-q) in screening for eating disorders in community samples, Behaviour research and therapy 42 (2004) 551–567.
[8] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Springer International Publishing, Thessaloniki, Greece, 2023.
[9] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, CoRR abs/2002.10957 (2020). URL: https://arxiv.org/abs/2002.10957. arXiv:2002.10957.
[10] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas, M.-J. Meurs, Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks., in: Clef (working notes), 2021, pp. 1031–1045.
[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[12] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[13] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, Mentalbert: Publicly available pretrained language models for mental healthcare, arXiv preprint arXiv:2110.15621 (2021).