

Lightweight Methods for Early Risk Detection

Diego Maupomé¹, Thomas Soulas¹, Fanny Rancourt¹, Ghyslain Cantin-Savoie¹,
Grégoire Winterstein¹, Sébastien Mosser³ and Marie - Jean Meurs¹

¹Université du Québec à Montréal, QC, Canada

³McMaster University, ON, Canada

Abstract

This paper describes the participation of the RELAI team in the eRisk 2023 shared tasks. The first task is a new problem introduced this year and consists of ranking sentences from a collection of user writings according to their relevance to a depression symptom. The second task is a recurring task, which highlights the problems of preliminary detection of gambling addiction.

Keywords

Mental Health, Natural Language Processing, Topic Modeling, Stylometry, BM25, Transformers, Domain adaptation

1. Introduction


The eRisk shared tasks seek to explore the application of Natural Language Processing (NLP) methods to estimate risk to mental health from online content. Its 2023 edition [1] included a task centered around finding excerpts related to signs and symptoms of depression, as well as a task aimed at the early detection of signs of pathological gambling. The present paper describes our participation to these tasks. Task 1, Search for symptoms of depression, was a retrieval task, aimed at finding and ranking sentences relevant to each of the signs and symptoms described in the Beck Depression Inventory, 2nd edition (BDI-II) [2]. The approach presented operates on the textual similarity between the sentences at hand and the questionnaire. Task 2, Early Detection of Signs of Pathological Gambling, is aimed at producing a singular assessment of the risk of pathological gambling given a subject's history of writings. This task proceeds iteratively, parsing a new writing per subject and expecting an update on the assessment for that individual. These assessments comprise a binary decision as well as a continuous score. Our participation was based on lightweight approaches deployed in past editions, such as stylometry and topic extraction, slightly modified to fit the context of eRisk 2023.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

✉ maupome.diego@courrier.uqam.ca (D. Maupomé); soulas.thomas_david@courrier.uqam.ca (T. Soulas);
rancourt.fanny.2@courrier.uqam.ca (F. Rancourt); cantin-savoie.ghyslain@courrier.uqam.ca (G. Cantin-Savoie);
winterstein.gregoire@uqam.ca (G. Winterstein); mossers@mcmaster.ca (S. Mosser); meurs.marie-jean@uqam.ca
(M. - J. Meurs)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Task 1: Search for symptoms of depression

2.1. Task and Data

The data consist of a set of independent sentences gathered from Reddit¹. No training set was provided. The testing dataset consists of 3,807,115 sentences from 3107 different Reddit users, averaging 1225.34 sentences per user. Each sentence contains an average of 13.63 words. This dataset is based on past eRisk data. The task is to select the sentences most relevant to a given item in the BDI and rank them accordingly. While the pool of sentences was shared across items, a separate ranking of up to 1000 sentences was to be provided per each of the 21 BDI items. In addition, each sentence in the ranking had to be attributed a numeric relevance score. This score had no preset scale; it should only be increasing with respect to relevance: sentences higher up in the ranking would have a higher score.

The question of what it means to be *relevant* remained largely underspecified in the description of the task, though annotation guidelines characterized relevant sentences as those conveying some information about the mental status of the author with respect to the topic discussed in the BDI item. Given the nature of the task, the relation of relevance can only be understood as one between a single discourse segment, usually a sentence, and a given item in the BDI, formalized as a question. Crucially, context plays no role in the relation at hand.

Yet, in spite of their differences, theoretical accounts of discursive relevance all highlight the highly contextual nature of relevance, especially when it is understood as the relationship between the elements in a question/answer pair (see a.o. Relevance Theory [3], accounts based on the notion of *Question under Discussion* [4], Segmented Discourse Representation Theory [5, 6], or Bayesian models of communication [7], all of them taking their roots in Grice's original program [8] and its notion of conversational implicature as the contextual enrichment of the content of an utterance). These theories emphasize that the meaning of a discourse is more than the sum of the meaning of its parts, and that context matters not just for the interpretation of the *content* of an utterance (e.g. by resolving indexical expressions or ambiguities), but also of the *use* of the utterance at a specific point in conversation, which in turns affects the interpretation of the content of the utterance. For example, in argumentative discourse, though a given proposition might not warrant, or even suggest, a particular conclusion, when taken in the larger context of an argument the same content might turn out to be very relevant to the conclusion because of its interaction with other discursive elements. This means that the setup of the task increases the risk of false negatives, i.e. of attributing a low relevance to sentence to a BDI item simply because we lack the necessary context to understand the speaker's point when using their sentence, and how it fits in their overall discourse. This is not to say that the task is inherently flawed: as already pointed out, early detection necessarily relies on limited data, and we approached the task in this spirit.

2.2. Approaches and Training

The approaches put forth for this task were centered around the relevance score, aiming to produce one score per sentence per item. From there, sentences were selected and ranked

¹<https://reddit.com>

based on this score. Each subject-produced sentence was matched against the sentences associated with each of the answer options of BDI items. Different off-the-shelf textual similarity models examined the resulting sentence pairs, yielding a similarity score. Then, the overall relevance score for a sentence with respect to an item was computed from its similarity scores associated with individual answers. Different manners of aggregating these similarity scores into a relevance score were considered, reflecting two different interpretations of relevance. The first of these interpretations—corresponding more closely to the one that would be given to annotators—is broader, characterizing relevance as touching on the same topics or subject matter as the pertinent BDI item. This interpretation was modeled by three separate aggregation operations: mean score, computing the relevance as the arithmetic mean of similarity scores, and minimum or maximum score, computing the relevance score as the lowest (resp. highest) similarity score. The second, more instrumentalist interpretation, conceived the relevance of a sentence as helping to select a specific answer to the BDI item for its author. Thus, a relevant sentence should display more clear affinity to one or a few answer options than the rest. This was modeled by reinterpreting similarity scores as logits and computing the entropy of their distribution. A relevant sentence would then have low entropy,² because it bears more distinct similarity to few answers.

Sentence similarities were computed using Transformer-based sentence encoders [9]. These were further adapted to the domain of discourse by training on past eRisk data [10]. However, given the large volume of sentences and the complexity of these models, encoding the entire set in order to compute similarities proved impracticable. To remediate this, sentences were first filtered by less computationally expensive BM25 models. Sentences bearing the highest similarity to some BDI sentence were retained and processed by Transformer-based models. Filtered out models were then assigned the lowest similarity found by the end model.

The BM25 model selected is a variant of Okapi BM25 [11]. As preprocessing, the sentences were fed through a Porter Stemmer followed by a Wordnet lemmatizer. Following [12], we set the minimal word length to 4. Stopwords were kept [11]. The top 100 sentences per BDI answer were retained for the Transformer encoder. Two related pretrained Transformer models were selected: a general purpose sentence Transformer (mpnetbase) and a semantic search-oriented variant (mpnetqa) [9]. As a means of domain adaptation, these were further trained on the SimCSE objective [13] with past eRisk data [14]. An additional, baseline run makes use only of the BM25 component. These approaches were evaluated by proxy using past eRisk data [10] under the interpretation of relevance as helping to select a specific answer to a given item [15], where mean-based aggregation produced the best results.

2.3. Results and Discussion

Rankings were evaluated using manual relevance assessment of a pool of sentences, extracted by a top-50 rule from participant submissions. Relevance judgments were combined using, separately, a majority and a unanimity rule. This results in sets of 50 to 350 relevant sentences per item for majority pooling and 20 to 260 for unanimity. Rankings were then evaluated against

²In practice, because the relevance score should be increasing, it was not computed as the entropy of the similarity score distribution, but as its Kullback-Liebler divergence with respect to the maximum entropy distribution with the same support (uniform).

these sets using standard metrics: Average Precision (AP), R-Precision (R-PREC), Precision at 10 (P@10) and NDCG at 1000 (NDCG@1000). Results are presented in Tables 1 for majority pooling and 2 for unanimity.

All our models exhibit poor performance. Across models, results deteriorate for unanimous relevance judgment, where relevant sentences are scarcer. Transformer models outperformed the baseline BM25 on all metrics but NDCG@1000. This may be due to the reconciliation of filtered sentences. When considering a larger number of sentences (*e.g.* 1000) more sentences with reconciled are likely to be present. Their equal similarities result in spurious relative order. More interestingly, continuing pretraining using the SimCSE objective hurts performance for both pretrained models. Further investigation is required to understand this deterioration. Domain-adapted models appear to produce lower, less varied relevance scores than their counterparts (mpnetbase: max: 0.59, std: 0.07, mpnetbase_simcse: max: 0.48, std: 0.05 ; mpnetqa: max: 0.72, std: 0.07, mpnetqa_simcse: max: 0.50, std: 0.06). Their discrepancies, however, go beyond bias. Examining the agreement between these variants reveals that they share between 400 and 650 sentences across items for mpnetbase and between 400 and 600 for mpnetqa. However, there is weak correlation in how they rank these common sentences, with mpnetbase and its domain-adapted variant exhibiting a Spearman coefficient of 0.34 on average across items and 0.30 for mpnetqa. Ultimately, the most limiting factor of our approaches may be the BM25 filtering step. The baseline BM25 approach obtains poor results on its own ranking of 1000 sentences, which shares only between 50 and 200 sentences per item with each of mpnetbase and mpnetqa, suggesting it may be a poor proxy for them.

Team	Run	AP	R-PREC	P@10	NDCG@1000
RELA	BM25	0.016	0.061	0.043	0.145
RELA	bm25 mpnetbase	0.048	0.081	0.538	0.140
RELA	bm25 mpnetbase_simcse	0.030	0.066	0.390	0.114
RELA	bm25 mpnetqa	0.038	0.075	0.438	0.126
RELA	bm25 mpnetqa_simcse	0.027	0.063	0.376	0.109
Formula-ML	SentenceTransformers_0.25	0.319	0.375	0.861	0.596

Table 1

Ranking-based evaluation for Task 1 (majority voting) by our models and the best performing model on each metric

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

3.1. Task and Data

The training dataset was composed of the test subjects from the 2022 and 2021 gambling task, namely 2384 subjects, 164 of which writings were positive (6.9%) for 2021, and 2079 subjects, 81 of which writings were positive (3.9%) for 2022. As for the test dataset, it was composed of 2071 subjects of which only 103 writings were positive (5.0%). Overall, considering we cannot guarantee that writings said positive or negative is really what has been determined [16], this is extremely unbalanced.

Team	Run	AP	R-PREC	P@10	NDCG@1000
RELAI	BM25	0.012	0.036	0.019	0.135
RELAI	bm25 mpnetbase	0.039	0.069	0.343	0.124
RELAI	bm25 mpnetbase_simcse	0.026	0.059	0.243	0.103
RELAI	bm25 mpnetqa	0.030	0.065	0.290	0.109
RELAI	bm25 mpnetqa_simcse	0.023	0.052	0.262	0.097
Formula-ML	SentenceTransformers_0.25	0.268	0.360	0.709	0.615
Formula-ML	SentenceTransformers_0.1	0.293	0.350	0.685	0.611

Table 2

Ranking-based evaluation for Task 1 (unanimity) by our models and the best performing models on each metric

3.2. Approaches and Training

The best approach from the previous iteration of the task was based on Nearest Neighbor retrieval [17]. Thus, we decided to explore possibilities with lightweight approaches even though, admittedly, Transformers approaches fared quite well [18]. The first one is a Stylometry approach. Building upon the approaches we tested at eRisk 2021 [19], three of our approaches use Topic Modeling [20] (ETM-50, ETM-50T2022, ETM-300T2022). Our last approach is a random one.

Stylometry considers stylometric features as a representation of the writing histories of subjects. These features include character and word n-gram frequencies, word and sentence lengths and character class frequencies. From this representation, a multilayer perceptron is trained to produce a decision. The best features were selected using halving random hyperparameter search [21], increasing the length of writing histories at each iteration.

ETM-50T, ETM-50T2022, ETM-300T2022 are ensemble approaches using ETM to vectorize. Using the same approach as [12], ETM is trained for 50 (ETM-50T, ETM-50T2022) and 300 (ETM-300T2022) topics on eRisk 2021 data. These representations are fed to two separate decision models, which must be in (positive) agreement for a positive decision to be made for the textual production of a given subject. The first one is a multilayer perceptron with 300 neurons per hidden layer trained on eRisk 2021 (ETM-50T) or on eRisk 2021 and 2022 data (ETM-50T2022, ETM-300T2022). Since the dataset is unbalanced, the threshold used for the perceptron is determined based on empirical observations made while testing on the eRisk 2021 dataset, respectively .34, .34 and .5. The second one is based on the approach we tried at eRisk 2021 [19]: finding the minimal similarity distance to be considered at risk of pathological gambling by computing the Hellinger distance between a self-evaluation questionnaires³ composed of 20 questions, and 199 testimonials⁴. The greatest such distance is chosen as a threshold for classifying textual production as positive. That is, using testimonials as pseudo-examples of

³<http://gamontreal.ca/>

⁴<https://gamblershelp.com.au/>

writing histories from positive textual production, any history at test time found to be closer to the questionnaire than the farthest testimonial can be assigned a positive label.

Random is a baseline run based on the overall distributions of labels from previous iterations of the task. The severity of problem gambling symptoms is assumed to lie on some finite, continuous range. The scores for a given set of subjects is naively set to follow a uniform distribution such that the share of positive subjects lies beyond a given threshold. Then, the proportion of positive subjects in the 2021 and 2022 datasets can be used to infer the gamma distribution from which they emerge (with large error given the small sample size). Sampling this distribution produces a uniform distribution which was used to attribute scores to subjects at each round.

3.3. Results and Discussion

Details about the metrics used on tables 3 and 4 can be found in [16, 14, 22]. While extremely under performing comparing to others approaches, two of our runs yield better F score than our Random one. Training our multi layer perceptron on 2022 data does not seem to help in yielding better results. This is supposedly due to the sheer number of negative subjects comparing to positive one. A better approach might have been to only add the positive subjects of 2022 to our training dataset.

System	Run	precision	recall	F ₁	ERDE ₅	ERDE ₅₀	latency _{TP}	speed	F _{latency}
Stylometry	0	0.000	0.000	0.000	0.047	0.047			
ETM-50T	1	0.058	0.971	0.109	0.048	0.039	1.0	1.000	0.109
ETM-50T2022	2	0.058	0.971	0.109	0.048	0.039	1.0	1.000	0.109
ETM-300T2022	3	0.000	0.000	0.000	0.047	0.047			
Random	4	0.047	1.000	0.090	0.047	0.047			
BioNLP-IISERB	3	1.000	0.049	0.093	0.045	0.045	1.0	1.000	0.093
ELiRF-UPV	0	1.000	0.883	0.938	0.026	0.010	4.0	0.988	0.927
UMUTeam	0	0.086	1.000	0.158	0.039	0.029	2.0	0.996	0.157

Table 3

Decision-based evaluation for task 2 by our models and the best performing models on each metric

Two of our approaches got a F1-score of 0 in Table 3, which suggests that these approaches could not predict one instance of pathological gambling. Considering that the results obtained with ETM-50 and ETM-50T2022 are not encouraging either, ETM-300T2022 results are probably due to an overly high threshold (0.5 while other ETM approaches are at .34) rather than the addition of a dataset. Seemingly, results for stylometry suggest that stylometric features are not enough to determine whether the textual production is positive or negative. The P@10 at 1 writing Table 4 being higher for stylometry than the other approach is probably because its scores were slightly different but the prediction was the same. Unsurprisingly, the random run performed poorly and ETM-50T approaches yield a slightly better F1 score. Though, considering the recall is extremely high contrary to the precision it just means that these approaches are overly biased toward the positive.

System	Run	1 writing			100 writings		
		P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
Stylometry	0	0.30	0.25	0.08	0.00	0.00	0.02
ETM-50T	1	0.00	0.00	0.01	0.00	0.00	0.00
ETM-50T2022	2	0.00	0.00	0.02	0.00	0.00	0.01
ETM-300T2022	3	0.00	0.00	0.01	0.00	0.00	0.01
Random	4	0.10	0.06	0.02	0.00	0.00	0.04
ELiRF-UPV	0	1.00	1.00	0.59	1.00	1.00	0.91
SINAI	1	1.00	1.00	0.73	1.00	1.00	0.90

System	Run	500 writings			1000 writings		
		P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
Stylometry	0	0.00	0.00	0.06	0.00	0.00	0.00
ETM-50T	1	0.00	0.00	0.00	0.00	0.00	0.00
ETM-50T2022	2	0.00	0.00	0.02	0.00	0.00	0.00
ETM-300T2022	3	0.00	0.00	0.00	0.00	0.00	0.00
Random	4	0.00	0.00	0.02	0.00	0.00	0.00
ELiRF-UPV	0	1.00	1.00	0.95	1.00	1.00	0.94
SINAI	1	1.00	1.00	0.85	0.00	0.00	0.00

Table 4

Ranking-based evaluation (P@10; NDCG@10; NDCG@100) for task 2 by our models and the best performing models on each metric

4. Conclusion

The 2023 edition of the eRisk workshop introduced a new task (Task 1: Search for symptoms of depression) whose objective is to consider a number of standalone sentences and rank them according to their relevance to each of the signs and symptoms described in the BDI. While it is sensible to try to detect sentences germane to the aspects of depression cataloged by the BDI, such a framework is highly local and may fail to capture the global patterns found in the writings of an individual. The approach proposed made use of the textual similarity between the sentences to rank and those found in the BDI items. Given the quantity of documents, this approach made use of two similarity models, with a lightweight BM25 model selecting the sentences to be examined by the more complex Transformer model. Future improvements could incorporate additional filtering steps with less complex models [23] and more sophisticated reconciliation of attrition. Task 2, Early Detection of Signs of Pathological Gambling, was in its third iteration. Our models appeared to suffer from poor calibration, yielding overtly similar scores, making decision policies difficult to establish.

Acknowledgments

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 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.
- [2] A. T. Beck, R. A. Steer, G. K. Brown, Beck Depression Inventory (BDI-II), *Psychological assessment* 10 (1996).
- [3] D. Sperber, D. Wilson, *Relevance: Communication and Cognition*, 2nd ed., Blackwell, Oxford, 1986.
- [4] C. Roberts, Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics, in: *OSU Working Papers in Linguistics*, volume 49: Papers in Semantics, Jae Haek Yoon and Andreas Kathol, 1996, pp. 91–136.
- [5] N. Asher, *Reference to abstract objects in discourse*, Kluwer, Dordrecht, 1993.
- [6] N. Asher, A. Lascarides, *Logics of Conversation*, Cambridge University Press, Cambridge, 2003.
- [7] A. Merin, Information, Relevance and Social Decision-Making, in: L. Moss, J. Ginzburg, M. de Rijke (Eds.), *Logic, Language, and computation*, volume 2, CSLI Publications, Stanford, 1999, pp. 179–221.
- [8] H. P. Grice, *Studies in the Way of Words*, Harvard University Press, Cambridge, 1989.
- [9] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Curran Associates Inc., Red Hook, NY, USA, 2020.
- [10] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2022: Pathological gambling, depression, and eating disorder challenges, in: *Advances in Information Retrieval*, Springer International Publishing, Cham, 2022, pp. 436–442.
- [11] A. Trotman, A. Puurula, B. Burgess, Improvements to bm25 and language models examined, in: *Proceedings of the 19th Australasian Document Computing Symposium, ADCS '14*, Association for Computing Machinery, New York, NY, USA, 2014, p. 58–65. doi:10.1145/2682862.2682863.
- [12] M. D. Armstrong, D. Maupomé, M.-J. Meurs, Topic models for assessment of mental health issues, in: *Proceedings of the Canadian Conference on Artificial Intelligence*, 2021. doi:10.21428/594757db.27574943.
- [13] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. URL: <https://aclanthology.org/2021.emnlp-main.552>. doi:10.18653/v1/2021.emnlp-main.552.
- [14] D. E. Losada, F. Crestani, J. Parapar, eRisk 2020: Self-harm and depression challenges, in: *Advances in Information Retrieval*, Springer International Publishing, Cham, 2020, pp. 557–563.
- [15] D. Maupomé, M. D. Armstrong, J. Alezot, R. Balassiano, M. Queudot, S. Mosser, M.-J. Meurs, Early mental health risk assessment through writing styles, topics and neural models, in:

- CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696, CEUR, 2020.
- [16] D. E. Losada, F. A. Crestani, A test collection for research on depression and language use, in: Conference and Labs of the Evaluation Forum, 2016.
 - [17] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, Uned-nlp at erisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors, in: Conference and Labs of the Evaluation Forum, 2022.
 - [18] A. M. Mármol-Romero, S. M. Jiménez-Zafra, F. M. Plaza-del-Arco, M. D. Molina-González, M.-T. Martín-Valdivia, A. Montejo-Ráez, SINAI at eRisk@CLEF 2022: Approaching Early Detection of Gambling and Eating Disorders with Natural Language Processing (2022).
 - [19] 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.
 - [20] A. B. Dieng, F. J. R. Ruiz, D. M. Blei, Topic Modeling in Embedding Spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453. doi:10.1162/tac1_a_00325.
 - [21] K. Jamieson, A. Talwalkar, Non-stochastic best arm identification and hyperparameter optimization, in: Artificial intelligence and statistics, PMLR, 2016, pp. 240–248.
 - [22] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 340–357.
 - [23] D. Maupomé, M.-J. Meurs, Contextualizer: Connecting the dots of context with second-order attention, Information 13 (2022) 290.