Awakened at eRisk 2026: Sentence Embedding-Based ADHD Symptom Ranking from Reddit Posts

Notebook for the eRisk Lab at CLEF 2026

Cosmin-Gabriel Becheru[1], Ciprian-Octavian Truică[1,2,*] and Elena-Simona Apostol[1]

[1]National University of Science and Technology POLITEHNICA Bucharest, Splaiul Independenţei 313, Bucureşti 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

This paper describes team Awakened's participation in Task 3 of the eRisk 2026 Lab at CLEF 2026. The task focuses on ranking Reddit sentences according to how relevant they are to the 18 ADHD symptoms listed in the Adult ADHD Self-Report Scale (ASRS-v1.1). We use an information retrieval pipeline based on pre-trained sentence embedding models. These models are used to compare ADHD symptom queries with a large set of Reddit sentences through semantic similarity. We submitted two runs based on different embedding choices. Run 1 uses all-MiniLM-L6-v2 and encodes only the sentence text, so each sentence is treated separately. Run 2 uses BAAI/bge-base-en-v1.5 and adds context by including the previous and next sentences in the input. For both runs, documents are ranked using cosine similarity. The results are then written in TREC format, with 1,000 retrieved sentences for each symptom. The paper explains how the corpus was parsed, how the sentence embeddings and rankings were produced, and how chunked GPU computation was used to process a corpus with millions of sentences on consumer hardware. It also compares the two retrieval settings and discusses possible ways to improve ranking at the symptom level.

Keywords

ADHD symptom detection, sentence embeddings, information retrieval, Reddit

1. Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most prevalent neurodevelopmental conditions, with Polanczyk et al. [1] reporting a worldwide-pooled prevalence of approximately 5% in children and adolescents and Simon et al. [2] estimating an adult prevalence of approximately 2.5%. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [3] defines ADHD through 18 symptom criteria organized into two domains: inattention (9 symptoms) and hyperactivity-impulsivity (9 symptoms). The Adult ADHD Self-Report Scale (ASRS-v1.1), developed by the World Health Organization in collaboration with Harvard researchers, operationalizes 18 corresponding items for adult self-screening and is the scale used by the eRisk 2026 organizers [4, 5]. Identifying these symptoms in naturalistic text, such as social media posts in which people describe everyday experiences, offers useful material for computational approaches to mental health assessment, but it also introduces several methodological difficulties.

The eRisk Lab at CLEF has gradually broadened its focus, moving from depression and anorexia detection toward a wider set of mental health conditions [6, 7, 8]. In the 2026 edition, Task 3 introduces the problem of ranking sentences from a large Reddit corpus according to their relevance to each of the 18 ASRS-v1.1 ADHD symptoms [4, 5]. Unlike classification tasks that assign binary labels, this ranking setting requires systems to generate more detailed relevance orderings, bringing the task closer to a traditional information retrieval problem. The task presents several difficulties. The corpus includes millions of Reddit sentences, which means that the system needs embedding and retrieval methods

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
cosmin.becheru@upb.ro (C. Becheru); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)
https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)
0009-0002-7720-8048 (C. Becheru); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)


that can run within the memory and computational limits of consumer-grade hardware. The 18 ADHD symptoms also vary in how specific they are. Some refer to concrete behavioral signs, such as fidgeting and losing things, while others describe more abstract experiences, such as being driven by a motor or having difficulty with quiet leisure activities. Therefore, the retrieval model needs to account for different degrees of semantic abstraction. In addition, individual sentences may express ADHD-related experiences without using clinical vocabulary, so the model must identify semantic similarity beyond direct keyword matching.

In this paper, we present our approach for Task 3 under the team name **Awakened**. We frame the problem as a dense retrieval task: each ADHD symptom is represented as a textual query, and each Reddit sentence is encoded into a dense vector using a pre-trained sentence embedding model. Ranking is performed via cosine similarity between query and document embeddings. We submitted two runs that differ in the embedding model and the text representation strategy, as detailed in Section 4.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 describes the task definition and data; Section 4 presents our methodology; Section 5 reports results; and Section 6 concludes with future directions.

2. Related Work

2.1. NLP Approaches to ADHD Detection

While depression has received the most attention in computational mental health research [9, 10], ADHD has been studied to a lesser extent in the NLP literature. Existing computational work tends to either operate on highly structured rating-scale data, such as the ASRS-v1.1 used in the present task, or attempt classification of ADHD-related discourse on a per-document level. The eRisk 2026 Task 3 represents one of the first large-scale benchmarks specifically targeting ADHD symptom identification in social media text at the sentence level [4, 5].

2.2. Sentence Embedding Models for Information Retrieval

Dense retrieval methods based on pre-trained sentence encoders have gained widespread adoption for semantic search and ranking tasks. Reimers and Gurevych [11] introduced Sentence-BERT (SBERT), which fine-tunes BERT-based models using siamese and triplet network structures to produce semantically meaningful sentence embeddings. The all-MiniLM-L6-v2 model, a distilled version of this approach, provides a practical balance between embedding quality and computational efficiency.

More recently, the BGE (BAAI General Embedding) family of models [12] has reported strong results on several retrieval benchmarks. BGE-base-en-v1.5, which we use in the context-aware run, was trained for retrieval tasks with contrastive learning and hard negatives, making it well-suited to the symptom ranking setting. Sentence-level embeddings combined with cosine similarity have also been used effectively in other NLP tasks. Truică et al. [13] showed that sentence transformers such as BERT and RoBERTa, together with cosine similarity scoring, can provide robust text representations for lexical simplification. This supports the broader use of the embedding-and-ranking approach adopted here for symptom retrieval.

2.3. Context-Aware Document Representation

A key question in sentence-level retrieval is whether incorporating surrounding context improves relevance ranking. Prior work in passage retrieval has shown that prepending or appending neighboring sentences can help disambiguate short or context-dependent texts. Our Run 2 explores this hypothesis by constructing context-enriched representations that include the preceding and following sentences alongside the target sentence.

3. Task Definition and Data

Task 3 of eRisk 2026 [4, 5] requires participants to rank Reddit sentences by their relevance to 18 ADHD symptoms. The symptoms are taken from the ASRS-v1.1 (which is itself aligned with the DSM-IV-TR and consistent with DSM-5 [3]) and cover both the inattention domain (Symptoms 1–9: careless mistakes, difficulty sustaining attention, difficulty listening, not finishing tasks, difficulty organizing, avoiding effortful tasks, losing things, easily distracted, forgetful in daily activities) and the hyperactivity-impulsivity domain (Symptoms 10–18: fidgeting, leaving seat, restlessness, difficulty with quiet leisure, driven by a motor, talking excessively, blurting out answers, difficulty waiting turn, interrupting or intruding).

The corpus follows the TREC format and contains Reddit sentences extracted from posts related to ADHD discussions. Each document is identified by a unique ID and contains the sentence text along with optional contextual fields (preceding and following sentences). For each of the 18 symptoms, participants must produce a ranked list of the top 1,000 most relevant documents.

The evaluation is based on two sets of relevance judgments, or qrels. In the majority voting setting, a document is treated as relevant when most annotators marked it as positive for the corresponding symptom. In the unanimity setting, a document is treated as relevant only when all annotators agreed that it was relevant.

Standard information retrieval metrics are used for evaluation, including Average Precision (AP), R-Precision, P@10, and NDCG@1000.

4. Methodology

Our approach has three stages. First, the corpus is parsed, with optional construction of sentence context. Second, each sentence is encoded using a pre-trained embedding model. Third, the sentences are ranked with cosine similarity, using GPU-accelerated chunked computation.

4.1. Corpus Parsing

The TREC-format corpus is parsed into a JSONL file where each line represents a single document with the following fields: `id` (document identifier), `text` (the target sentence), `pre` (preceding sentence, if available), and `post` (following sentence, if available). Square brackets and escape characters are sanitized to prevent tokenization artifacts.

4.2. Run 1: Text-Only Embedding with MiniLM

In Run 1, each document is encoded using only its `text` field with the all-MiniLM-L6-v2 model [11]. This lightweight model generates 384-dimensional sentence embeddings and supports fast inference, which makes it practical for encoding large corpora. The embeddings are L2-normalized so that cosine similarity can be computed through the dot product.

4.3. Run 2: Context-Aware Embedding with BGE-base

Run 2 addresses a limitation of text-only encoding, since short or ambiguous sentences may not contain enough information for accurate symptom matching. Each document is encoded using a context-enriched representation built by concatenating the available context fields:

```
Previous sentence: <pre>
Current sentence: <text>
Next sentence: <post>
```

This enriched text is encoded using BAAI/bge-base-en-v1.5 [12], a 768-dimensional embedding model specifically optimized for retrieval tasks. The model was loaded in float16 precision to reduce memory consumption on GPU.

4.4. Symptom Query Construction

Each of the 18 ADHD symptoms is represented as a natural language query that expands the symptom name with descriptive phrases covering the behavioral manifestations listed in the ASRS-v1.1 (and DSM-5-equivalent) criteria. For example, Symptom 1 ("Careless mistakes") is expanded to: *"Making careless mistakes, missing details, overlooking parts of tasks, not paying close attention during boring or repetitive work."*

These symptom queries are embedded using the same model as the corpus documents, ensuring that query and document vectors reside in the same embedding space.

4.5. Ranking via Cosine Similarity

For each symptom query, the ranking score of a document is computed as the dot product between the L2-normalized query embedding and the L2-normalized document embedding (equivalent to cosine similarity). The top 1,000 documents per symptom are retained.

4.6. GPU-Accelerated Chunked Computation

Given the scale of the corpus, which contains millions of documents, computing all similarity scores in a single operation would exceed the available GPU memory. For this reason, we use a chunked computation strategy. The document embedding matrix is divided into chunks of 500,000 rows. For each chunk, similarity scores against the query are computed on GPU, using CUDA or MPS, through a single matrix-vector multiplication. The top-K scores from each chunk are kept. After all chunks have been processed, the top-K results from each chunk are merged and sorted globally to obtain the final top-1,000 ranking. This strategy makes it possible to process very large corpora while keeping GPU memory usage constant, proportional to the selected chunk size. It also allows the similarity computation to use the GPU efficiently. After each chunk is processed, memory is explicitly released through garbage collection and GPU cache clearing.

4.7. Checkpointing and Resumability

To handle potential interruptions during the ranking of 18 symptoms across the large corpus, we implement a per-symptom checkpointing mechanism. Each symptom's results are saved to an individual temporary file upon completion. On restart, already-processed symptoms are skipped, allowing the pipeline to resume from the point of interruption. The final TREC run file is assembled by concatenating all per-symptom result files.

Table 1 summarizes the key differences between the two submitted runs.

Table 1
Comparison of the two submitted runs.

Aspect	Run 1	Run 2
Model	all-MiniLM-L6-v2	BAAI/bge-base-en-v1.5
Embedding dim.	384	768
Input text	Sentence only	Sentence + context
Precision	float32	float16
Top-K per symptom	1,000	1,000

5. Results and Discussion

5.1. Official Results

The official evaluation of Task 3 was carried out by the eRisk 2026 organizers using the two qrels files, majority-vote and unanimity, against the submitted TREC runs [4, 5]. Our two submissions, text-only MiniLM (Run 1) and context-aware BGE (Run 2), reflect two sentence embedding-based retrieval strategies for symptom ranking. Table 2 presents the official scores released by the organizers.

Table 2
Awakened's official Task 3 scores under both relevance-judgment settings [4, 5].

Qrels	Run	AP	R-PREC	P@10	NDCG@1000
Majority-vote	Run 1 (MiniLM)	0.006	0.015	0.022	0.057
Majority-vote	Run 2 (BGE+ctx)	0.006	0.019	0.028	0.058
Unanimity	Run 1 (MiniLM)	0.009	0.017	0.017	0.054
Unanimity	Run 2 (BGE+ctx)	0.005	0.011	0.017	0.043

The absolute scores are low, situating our runs in the lower portion of the official ranking and well below the leading systems (e.g., the top NeuroRankUB run reaches AP ≈ 0.248 under the majority-vote setting [4, 5]). This indicates that a vanilla bi-encoder pipeline without domain adaptation, query reformulation, or re-ranking is insufficient for this ranking task. The two runs are close to each other on every metric, with no consistent winner across qrels.

5.2. Internal Comparison of Runs

An internal comparison of the top-ranked documents between Run 1 and Run 2 reveals substantial differences in the retrieved document sets. A quantitative overlap analysis across all 18 symptoms shows that, on average, only a fraction of the top-10 results are shared between the two runs, confirming that the choice of embedding model and context representation has a meaningful impact on retrieval behavior.

Qualitative inspection of the top-ranked documents indicates that the context-aware BGE model (Run 2) often retrieves documents in which ADHD-relevant information appears across both the target sentence and its surrounding context. In contrast, the text-only MiniLM model (Run 1) more often retrieves documents where the symptom is stated directly in the target sentence itself.

5.3. Discussion

Model selection. The BGE-base-en-v1.5 model was chosen because it was trained for retrieval tasks using hard negative mining. For this reason, it may provide embeddings that are more useful for ranking than MiniLM, which is more oriented toward general semantic similarity. However, this choice also has a cost. BGE produces 768-dimensional embeddings, while MiniLM produces 384-dimensional embeddings, so the BGE representation needs about twice as much storage and computation. This makes the choice between the two models a practical balance between retrieval quality and available resources.

Context enrichment. In Run 2, surrounding sentences were added because Reddit users often describe ADHD-related experiences over several sentences rather than in one isolated sentence. In some cases, the target sentence alone is too short or too vague to be matched accurately to a symptom. For example, a sentence such as I just can't do it anymore is difficult to interpret by itself, but it becomes more informative when it follows Every time I sit down to do homework, I get distracted within minutes. The context-aware representation was therefore used to include information that appears across neighboring sentences.

Symptom heterogeneity. The 18 ADHD symptoms are not expressed in text in the same way. Some behavioral symptoms, such as fidgeting or losing things, are usually described through direct and

observable actions. Other symptoms, such as difficulty sustaining attention or being driven by a motor, may appear in more subjective or figurative descriptions. Because of this variation, the same embedding method may not be equally effective for every symptom. Future work could therefore test fine-tuning strategies adapted to individual symptoms.

Limitations. Both runs rely on general-purpose sentence embedding models and do not include fine-tuning on ADHD-related or broader mental health text. The cosine similarity ranking also ignores the position of a sentence within a longer post or thread, although this information could help estimate relevance. The symptom queries were manually derived from the ASRS-v1.1 item descriptions, and recall might be improved through data-driven query expansion, for example, by using terms found in ADHD-related discussions. Finally, the current method ranks each symptom separately and does not model possible relations between symptoms or comorbidity patterns.

6. Conclusions and Future Work

We described a sentence embedding approach for ranking ADHD symptoms in Reddit posts as part of eRisk 2026 Task 3. The method treats the task as dense retrieval: symptom queries and Reddit sentences are encoded in the same embedding space, and the final ranking is produced with cosine similarity. We tested two settings. The first one uses MiniLM and encodes only the sentence text. The second one uses BGE-base and adds the surrounding context to the sentence representation.

The pipeline is designed to work with very large collections of documents. It uses GPU-based dense retrieval with chunked computation, so the corpus can be processed without loading all embeddings into GPU memory at once. We also compared text-only encoding with context-aware encoding in the setting of mental health symptom retrieval. In addition, the system uses per-symptom checkpointing, which makes the ranking process easier to resume and more stable when working with large-scale data.

Future work could focus on several directions. One direction is to adapt the sentence embedding models to ADHD-related Reddit data, for example, by using contrastive learning with symptom-labeled positive pairs. Another direction is cross-encoder re-ranking, where a cross-encoder model would be used to re-rank the top-K candidates produced by the bi-encoder stage, with the aim of improving precision at low cutoffs [14]. Query expansion could also be explored by adding terms extracted from ADHD discussions or clinical literature to the symptom queries. A further extension would be to model several symptoms together, rather than treating each one separately, so that comorbidity patterns and dependencies between symptoms can be considered. Finally, dense retrieval could be combined with sparse retrieval methods, such as BM25, using reciprocal rank fusion to improve recall.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks, paraphrasing, and writing-style improvements, and ChatGPT for light language-editing support and formatting assistance. These tools were used solely to improve clarity, readability, and presentation. After using these tools, the authors carefully reviewed, edited, and verified the content as needed and take full responsibility for the publication's content.

References

[1] G. Polanczyk, M. S. de Lima, B. L. Horta, J. Biederman, L. A. Rohde, The worldwide prevalence of ADHD: A systematic review and metaregression analysis, American Journal of Psychiatry 164 (2007) 942–948. doi:10.1176/ajp.2007.164.6.942.

[2] V. Simon, P. Czobor, S. Bálint, A. Mészáros, I. Bitter, Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis, British Journal of Psychiatry 194 (2009) 204–211. doi:10.1192/bjp.bp.107.048827.

[3] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed., American Psychiatric Publishing, Arlington, VA, 2013.

[4] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026 Early Risk Prediction on the Internet: Symptom Ranking and Conversational Approaches for Depression and ADHD (Extended Overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[5] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026 Early Risk Prediction on the Internet: Symptom Ranking and Conversational Approaches for Depression and ADHD, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Lecture Notes in Computer Science, Springer, 2022, pp. 233–256. doi:10.1007/978-3-031-13643-6_18.

[7] 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, Thessaloniki, Greece, September 18–21, 2023, Proceedings, Lecture Notes in Computer Science, Springer, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[8] 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, Grenoble, France, September 9–12, 2024, Proceedings, Lecture Notes in Computer Science, Springer, 2024, pp. 73–92. doi:10.1007/978-3-031-71908-0_4.

[9] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: A narrative review, npj Digital Medicine 5 (2022) 46. doi:10.1038/s41746-022-00589-7.

[10] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, ACM, 2013, pp. 47–56. doi:10.1145/2464464.2464480.

[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[12] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-pack: Packed resources for general chinese embeddings (2024) 641–649. doi:10.1145/3626772.3657878.

[13] C.-O. Truică, A.-I. Stan, E.-S. Apostol, SimpLex: A lexical text simplification architecture, Neural Computing and Applications 35 (2023) 6265–6280. doi:10.1007/s00521-022-07905-y.

[14] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, University Politehnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.