AI-MH-Z at eRisk@ CLEF 2026: Ranking ADHD Symptom Sentences Using Weighted Semantic Embedding Strategies

Notebook for the eRisk Lab at CLEF 2026

Zarana Ramani[1,†], Dr. Hiteishi Diwanji[2,†]

[1]Research Scholar, Gujarat Technological University, Ahmedabad, India
[2]Professor, L.D. College of Engineering, Ahmedabad, India

Abstract

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most prevalent neurodevelopmental disorders, affecting both children and adults worldwide. Yet, it remains widely underdiagnosed due to the subtle nature of its symptoms and the reliance on self-reported assessments in clinical practice. This paper presents our approach for eRisk CLEF 2026 Task 3, a novel shared task focused on ranking social media sentences according to their relevance to the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1). Given the absence of annotated training data, we propose an unsupervised dense retrieval approach in which candidate sentences and enriched symptom descriptors are encoded using the all-MiniLM-L6-v2 sentence transformer and indexed in a Qdrant vector database for efficient similarity-based retrieval. We submitted four runs employing distinct weighting strategies: uniform option weighting, ordinal severity weighting, custom per-symptom weighting, and direct symptom-description similarity. Experimental results under both the majority-vote and unanimity evaluation schemes consistently show that option-level similarity aggregation outperforms direct symptom-description matching, with the best option-weighted run achieving an NDCG of 0.317 and 0.291 under the majority-vote and unanimity schemes, respectively, compared to 0.150 and 0.129 for the description-based run.

Keywords

ADHD Detection, Sentence Transformers, Dense Retrieval, Sentence Ranking, Natural Language Processing (NLP), Weighted Similarity

1. Introduction

Adult Attention Deficit Hyperactivity Disorder (ADHD) is a mental health disorder that affects how the brain regulates attention, impulse control, and activity levels [1]. It is one of the most commonly diagnosed neurodevelopmental conditions in children, with symptoms persisting into adulthood in a significant proportion of cases [2]. In recent years, social media platforms have become an integral part of daily life and a valuable source for understanding individuals' mental health experiences. People openly discuss ADHD-related challenges, including difficulties with attention, impulsivity, and hyperactivity, on platforms such as Reddit and Twitter, generating large volumes of informal yet clinically meaningful text.

This paper presents our approach to CLEF eRisk 2026 Task 3 [3, 4], a shared task aimed at ranking candidate sentences according to their relevance to the 18 symptoms defined in the Adult ADHD Self-Report Scale (ASRS-v1.1) [5]. We propose a system based on unsupervised dense retrieval using sentence-transformer embeddings, combined with multiple option-level weighting strategies to identify the top 1,000 candidate sentences for each symptom. We evaluate four distinct scoring configurations, analyze their performance under two relevance judgment schemes, and discuss potential directions for improving ADHD symptom detection from social media text.

The structure of this paper is as follows: Section 2 reviews related work. Section 3 describes the task and the dataset. Section 4 presents the proposed methodology. Section 5 discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future research.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
zaranaramani@gmail.com (Z. Ramani); hiteishi@ldce.ac.in (Dr. H. Diwanji)

2. Related Work

Early detection of mental health conditions from social media text has received increasing attention over the past decade. This trend is driven by the growing availability of user-generated content and advances in natural language processing. Transformer-based language models have demonstrated remarkable effectiveness across a wide range of NLP tasks in the healthcare domain. Recent surveys highlight their growing use for the text-based identification of mental health conditions, including depression, anxiety, and ADHD [6].

Early work in this area focused primarily on the binary classification of ADHD-related content rather than fine-grained symptom-level analysis. Alsharif et al. [7] developed a diagnostic system for ADHD by applying machine learning and deep learning techniques, including gated recurrent units and long short-term memory networks, to Reddit data. Their results demonstrated that textual features extracted from social media posts can reliably identify ADHD-related content. However, their approach treated ADHD as a single category without distinguishing between individual symptom dimensions, limiting its clinical interpretability. Similarly, Karim et al. [8] applied knowledge distillation to develop a lightweight BERT-based model for classifying the severity of ADHD-related concerns expressed in social media text, demonstrating that compact transformer models can effectively capture nuanced symptom expressions. Although this work moved toward severity-level analysis, it still operated at the post level rather than the individual sentence level.

At a broader level, Kang et al. [9] conducted a large-scale linguistic and topic analysis of ADHD discussions on Reddit spanning more than a decade, providing evidence that users express ADHD-related experiences through consistent and identifiable language patterns. Although this work demonstrates the richness of social media text for ADHD research, it focused on population-level trends rather than individual symptom relevance, leaving the problem of fine-grained symptom-to-sentence alignment unexplored. Likewise, Barrios et al. [10] demonstrated that NLP and stylometric analysis can enhance ADHD diagnosis by providing objective linguistic markers, although they noted that further research is needed to validate these methods on larger and more diverse populations. However, their study was limited to structured narrative texts written by adolescents, leaving open the challenge of applying similar methods to informal social media text produced by adults.

Symptom-level sentence ranking from social media was first introduced at eRisk 2023, where participants ranked sentences according to their relevance to the 21 BDI-II depression symptoms, with the top-performing systems relying on transformer-based embeddings and cosine similarity [11]. Subsequent editions explored Sentence-BERT with query expansion [12] and option-level weighted similarity aggregation methods, including weighted aggregation across response options [13]. The latter demonstrated that decomposing symptoms into individual response options improves retrieval performance. However, all previous eRisk sentence-ranking tasks have focused exclusively on depression, leaving ADHD symptom-level sentence ranking unexplored. Our work addresses this gap by adopting a similar option-level retrieval framework while applying it to ADHD in a fully unsupervised setting without annotated training data.

3. Task Definition and Dataset description

3.1. ADHD Symptom Sentence Ranking

In this work, we address Task 3 of eRisk 2026, entitled *ADHD Symptom Sentence Ranking*. The objective of this task is to rank sentences from social media posts according to their relevance to each of the 18 symptoms defined in the Adult ADHD Self-Report Scale (ASRS-v1.1). A sentence is considered relevant if it contains any indication of the user's condition with respect to a particular ADHD symptom, regardless of its polarity or stance.

3.2. Dataset Details

The official TREC-formatted corpus was provided as part of the eRisk 2026 shared tasks [3, 4]. It consists of sentence-level social media posts collected from publicly available sources to capture ADHD-related expressions. The corpus contains 4,170,875 candidate sentences, with an average sentence length of 13.05 words.

Each entry in the TREC collection contains the following four fields:

- <DOCNO>: a unique document identifier;
- <PRE>: the preceding context;
- <TEXT>: the target sentence;
- <POST>: the following context.

3.3. Evaluation Metrics

System performance is evaluated using four standard information retrieval metrics. Mean Average Precision (MAP) measures the average precision across all recall levels, rewarding systems that rank relevant sentences higher overall. R-Precision (R-PREC) measures precision at the rank position equal to the total number of relevant sentences for a given symptom. Precision at rank 10 (P@10) measures the proportion of relevant sentences among the top 10 retrieved results, providing a user-oriented assessment of ranking quality. Finally, Normalized Discounted Cumulative Gain (NDCG) evaluates the overall ranking quality by assigning greater importance to relevant sentences appearing near the top of the ranked list while penalizing those ranked lower. Relevance judgments are obtained through top-k pooling across all participating systems, followed by expert human annotation under two evaluation schemes: majority-vote, where a sentence is considered relevant if at least two of the three assessors agree, and unanimity, where all three assessors must judge the sentence as relevant.

4. Proposed Methodology

This section describes the proposed methodology for ranking candidate sentences according to their relevance to each of the 18 ADHD symptoms defined in the ASRS-v1.1 questionnaire. The overall framework consists of four stages: data extraction and preprocessing, ADHD symptom descriptors, semantic embedding, and relevance ranking.

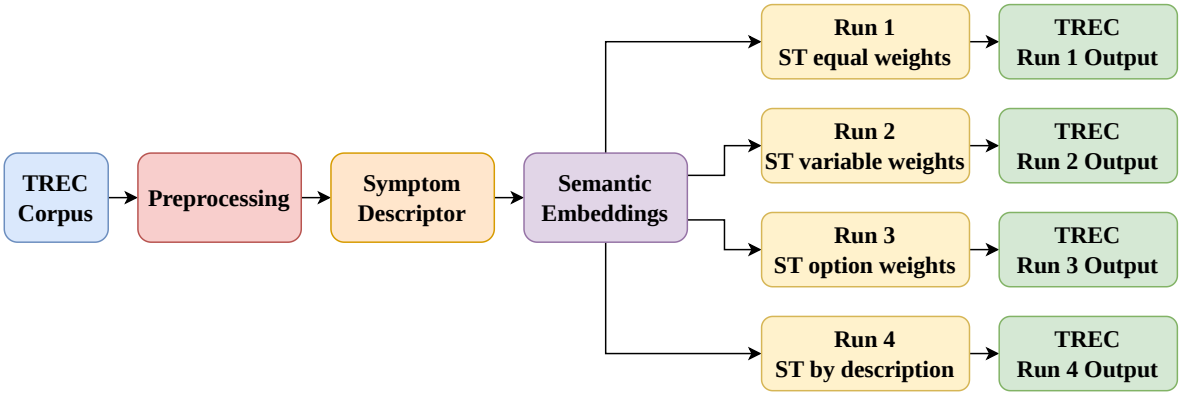

Figure 1: Overview of the proposed system framework for ADHD symptom sentence ranking

4.1. Data Extraction and Preprocessing

Candidate sentences were extracted from the provided TREC-formatted corpus containing 4,170,875 sentences. Each document was parsed using regular-expression-based extraction to retrieve the <DOCNO>,

<PRE>, <TEXT>, and <POST> fields. Only documents containing valid <DOCNO> and <TEXT> fields were retained.

After extraction, text normalization was applied independently to the <PRE>, <TEXT>, and <POST> fields. Records containing missing or null values were removed. The remaining text was converted to lowercase, non-essential characters were removed, and extra whitespace was normalized. Finally, the cleaned corpus was stored in CSV format for the subsequent embedding and indexing stages.

4.2. ADHD Symptom Descriptors

The ASRS-v1.1 questionnaire consists of 18 symptom items, each expressed as a question describing a specific ADHD-related behavioral pattern. Rather than using the original questions directly as search queries, each symptom was manually rewritten as a concise descriptive phrase that captures its core meaning in plain language. For example, the symptom question *"How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?"* was reformulated as *"difficulty finishing tasks, trouble completing final details, not wrapping up projects"*, providing a compact representation of the key behavioral signals that relevant sentences may express.

In addition to the symptom-level descriptors, each of the five Likert-scale response options was rewritten as a complete natural-language sentence reflecting the user's experience at the corresponding frequency level. For the same symptom, the response options ranging from *Never* to *Very Often* were expanded into complete first-person statements, such as *"I never have trouble wrapping up the final details of a project"* and *"I very often have trouble finishing the final details of a project once the challenging parts are done"*.

The same procedure was applied consistently across all 18 symptoms.

4.3. Semantic Embedding

For each document, a single sentence representation was constructed by concatenating the <PRE>, <TEXT>, and <POST> fields. All candidate sentences were then encoded into dense vector representations using the ***all-MiniLM-L6-v2*** sentence-transformer model [14, 15]. This model produces semantically meaningful 384-dimensional embeddings while providing an effective balance between retrieval accuracy and computational efficiency, making it well suited for large-scale semantic search. The resulting embeddings were indexed in a Qdrant vector database to enable efficient approximate nearest-neighbor retrieval across the entire corpus.

Similarly, the enriched symptom descriptors and their corresponding response-option statements were encoded using the same embedding model to ensure that both queries and candidate sentences were represented within a common semantic embedding space.

4.4. Relevance Ranking

To investigate different sentence-ranking strategies, four experimental runs were designed, each employing a different weighting scheme for computing the final relevance score.

- **Run 1 — ST equal weights - Uniform Option Weighting**
  In this baseline configuration, cosine similarity was computed between each candidate sentence and the five response-option representations associated with a given symptom. Equal weights of 0.2 were assigned to each response option, treating every frequency level as equally informative. The final relevance score was calculated as the weighted sum of the five cosine similarity scores, providing a simple and unbiased baseline for comparison.
- **Run 2 — ST variable weights - Ordinal Severity Weighting**
  This run introduced differential weights based on the ordinal position of each response option, reflecting the increasing diagnostic significance of higher-frequency responses. Weights were assigned as $[0.05, 0.10, 0.20, 0.30, 0.35]$ for the five options from "Never" to "Very Often" respectively, placing greater emphasis on sentences that align with more frequent and severe

manifestations of the symptom. The final relevance score was a weighted aggregation of cosine similarities scaled by option severity level.
This configuration assigned progressively larger weights to higher-frequency response options to reflect their greater clinical significance. Specifically, weights of $[0.05, 0.10, 0.20, 0.30, 0.35]$ were assigned to the response options ranging from *Never* to *Very Often*. The final relevance score was obtained as the weighted sum of the corresponding cosine similarity scores.

- **Run 3 — ST option weights - Custom Symptom-Specific Weighting**
  Instead of applying a single weighting scheme to all symptoms, this configuration assigned symptom-specific weights based on the shaded scoring regions defined in the official ASRS-v1.1 scoring rubric. In the questionnaire, particular response options are highlighted with darker shading to indicate clinically significant responses that contribute more strongly to ADHD screening. Accordingly, higher weights were assigned to these response options, while lower weights were assigned to the remaining options.
  For symptoms in which the shaded region includes the *Sometimes*, *Often*, and *Very Often* response options, the weights were set to $[0.125, 0.125, 0.25, 0.25, 0.25]$. For symptoms in which only the *Often* and *Very Often* options are shaded, the weights were set to $[0.10, 0.10, 0.10, 0.35, 0.35]$. In both cases, the weights sum to 1.0. The final relevance score for each candidate sentence was computed as the weighted sum of the cosine similarity scores between the sentence embedding and the five response-option embeddings. This weighting strategy directly reflects the clinical scoring logic of the ASRS-v1.1 by emphasizing response options that are more diagnostically informative.
- **Run 4 — ST by description - Symptom Description Similarity**
  Unlike the previous configurations, this approach did not decompose each symptom into individual response options. Instead, cosine similarity was computed directly between each candidate sentence and the corresponding enriched symptom descriptor described in Section 4.2. Consequently, sentences expressing the overall semantic meaning of a symptom were ranked higher regardless of the implied response frequency.

5. Results

The evaluation results presented in Tables 1 and 2 summarize the performance of the top three participating systems alongside our four submitted runs under two relevance judgment schemes. Although our system did not achieve top-ranked performance in the shared task, the submitted runs produced competitive retrieval results, demonstrating that dense semantic similarity using sentence-transformer embeddings combined with weighted option-level matching provides a viable approach for ADHD symptom sentence ranking.

Table 1
Ranking-Based Evaluation (majority-vote) — Top 3 Runs + AI-MH-Z Team

Rank	Team	Run	AP	R-PREC	P@10	NDCG
1	NeuroRankUB	final ADHD ranking	**0.248**	**0.328**	**0.528**	**0.547**
2	NeuroRankUB	gte rrf final	0.203	0.264	0.450	0.516
3	erisk-cedri	EnsembleV3	0.190	0.262	0.400	0.461
15	AI-MH-Z	ST equal weights	**0.076**	**0.110**	**0.239**	**0.317**
16	AI-MH-Z	ST option weights	0.075	**0.110**	0.228	0.316
17	AI-MH-Z	ST variable weights	0.074	0.109	0.228	0.313
31	AI-MH-Z	ST by description	0.035	0.062	0.094	0.150

Under the majority-vote scheme, the top-ranked system was NeuroRankUB with its *final ADHD ranking* run, achieving an AP of 0.248, R-PREC of 0.328, P@10 of 0.528, and an NDCG of 0.547. Among our submissions, the *ST equal weights* run achieved the best performance, with an AP of 0.076, an

Table 2
Ranking-Based Evaluation (Unanimity) — Top 3 Runs + AI-MH-Z Team

Rank	Team	Run	AP	R-PREC	P@10	NDCG
1	NeuroRankUB	final ADHD ranking	**0.207**	**0.274**	**0.367**	**0.488**
2	NeuroRankUB	gte rrf final	0.184	0.233	0.306	0.466
3	NeuroRankUB	ADHD submission	0.166	0.232	0.311	0.439
14	AI-MH-Z	ST equal weights	**0.078**	**0.108**	**0.167**	**0.291**
15	AI-MH-Z	ST option weights	**0.078**	0.105	0.161	0.289
16	AI-MH-Z	ST variable weights	0.076	0.102	0.167	0.288
34	AI-MH-Z	ST by description	0.031	0.051	0.072	0.129

R-PREC of 0.110, a P@10 of 0.239, and an NDCG of 0.317, ranking 15th overall. The *ST option weights* and *ST variable weights* runs produced very similar results, suggesting that the different weighting strategies had only a marginal impact on retrieval performance under this evaluation scheme. In contrast, the *ST by description* run recorded the lowest scores across all metrics, achieving an NDCG of 0.150 and ranking 31st overall.

Under the stricter unanimity scheme, NeuroRankUB again occupied the top three positions, with its *final ADHD ranking*, *gte rrf final*, and *ADHD submission* runs achieving NDCG scores of 0.488, 0.466, and 0.439, respectively. This consistent performance across both evaluation schemes suggests that the team's retrieval approach is robust to differences in annotation criteria. Among our submissions, the *ST equal weights* run again achieved the best performance, with an AP of 0.078 and an NDCG of 0.291, ranking 14th overall. Interestingly, the *ST option weights* run achieved the same AP of 0.078 under this scheme, although its NDCG was slightly lower at 0.289. The *ST variable weights* run followed closely with an NDCG of 0.288. As in the majority-vote evaluation, the *ST by description* run produced the weakest performance, achieving an NDCG of 0.129 and ranking 34th overall.

Across both evaluation schemes, the results indicate that option-level similarity aggregation consistently outperforms direct symptom-description matching. Among the option-level approaches, the three weighting strategies produced very similar performance, suggesting that weighting has only a limited effect on retrieval effectiveness. Notably, the simple uniform weighting strategy yielded the best overall performance among our submissions, indicating that more complex weighting schemes did not provide a substantial advantage for this task. Furthermore, the consistent ordering of our runs across both evaluation schemes suggests that the proposed retrieval framework is relatively stable under different relevance judgment criteria.

6. Conclusion

Our results demonstrate that dense retrieval using sentence-transformer embeddings with option-level weighting provides a viable approach for ADHD symptom sentence ranking in a fully unsupervised setting. Across both evaluation schemes, the *ST equal weights*, *ST option weights*, and *ST variable weights* runs achieved very similar NDCG scores of 0.317, 0.316, and 0.313, respectively, under the majority-vote scheme, and 0.291, 0.289, and 0.288, respectively, under the unanimity scheme. These findings indicate that the choice of weighting strategy has only a limited impact on retrieval effectiveness, suggesting that the semantic information captured through option-level similarity is sufficiently informative regardless of how weights are distributed across response options. In contrast, the *ST by description* run achieved substantially lower NDCG scores of 0.150 under the majority-vote scheme and 0.129 under the unanimity scheme—a decrease of approximately 0.167 and 0.162 NDCG points, respectively, compared with the best-performing option-level approach. This result highlights that relying solely on a single symptom-level descriptor, without decomposing symptoms into individual response options, substantially reduces the system's ability to retrieve relevant sentences.

Future work will investigate more powerful embedding models, including domain-adapted trans-

formers such as MentalBERT and other models pre-trained on clinical and mental health corpora, which may better capture the linguistic characteristics of ADHD-related expressions. In addition, cross-encoder-based re-ranking techniques will be explored to refine the initial retrieval results, along with ensemble methods that combine multiple retrieval signals. Finally, incorporating richer clinical knowledge into the weighting strategy and investigating data augmentation techniques may further improve retrieval performance for ADHD symptom sentence ranking.

Declaration on Generative AI

During the preparation of this work, we used ChatGPT and Claude for grammar, spelling, and language refinement. After using these tools, we carefully reviewed and edited the content as needed and take full responsibility for the content of this publication.

References

[1] The Online Psychologists, Adhd, 2026. URL: https://www.theonlinepsychologists.co.uk/expertise/adhd/, accessed: 2026-05-25.

[2] G. Ayano, L. Tsegay, Y. Gizachew, M. Necho, K. Yohannes, S. Demelash, T. Anbesaw, R. Alati, Prevalence of attention deficit hyperactivity disorder in adults: Umbrella review of evidence generated across the globe, Psychiatry Research 330 (2023) 115578. doi:10.1016/j.psychres.2023.115449.

[3] 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.

[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] R. C. Kessler, L. Adler, M. Ames, O. Demler, S. Faraone, E. Hiripi, R. Jin, T. Spencer, T. B. Ustun, E. E. Walters, The World Health Organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population, Psychological Medicine 35 (2005) 245–256. doi:10.1017/S0033291704002892.

[6] C. M. Greco, A. Tagarelli, E. Zumpano, Transformer-based language models for mental health issues: A survey, Pattern Recognition Letters 167 (2023) 204–211. doi:10.1016/j.patrec.2023.02.016.

[7] N. Alsharif, et al., ADHD diagnosis using text features and predictive machine learning and deep learning algorithms, Journal of Disability Research 3 (2024). doi:10.57197/JDR-2024-0082.

[8] A. A. J. Karim, K. H. M. Asad, M. G. R. Alam, Larger models yield better results? streamlined severity classification of adhd-related concerns using bert-based knowledge distillation, PLOS ONE 20 (2025) e0315829. doi:10.1371/journal.pone.0315829.

[9] J. Kang, N. Haslam, M. Conway, Converging representations of ADHD and autism on social media: Linguistic and topic analysis of trends in Reddit data, Journal of Medical Internet Research 27 (2025) e70914. doi:10.2196/70914.

[10] M. Barrios, D. Poznyak, J. Y. Lee Samson, H. Rafi, S. Gabay, F. Cafiero, M. Debbané, Detecting ADHD through natural language processing and stylometric analysis of adolescent narratives, Frontiers in Child and Adolescent Psychiatry 4 (2025) 1519753. doi:10.3389/frcha.2025.1519753.

[11] D. E. Losada, J. Parapar, A. Barreiro, Overview of eRisk at CLEF 2023: Early risk prediction on the Internet, in: Proceedings of the CLEF 2023 Working Notes, CEUR Workshop Proceedings, 2023.

[12] Z. Z. Ang, et al., Semantic retrieval of BDI symptoms in user writings, in: Proceedings of the CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025.
[13] J. Bento, L. Coheur, INESC-ID at eRisk 2025: Exploring fine-tuned, similarity-based, and prompt-based approaches to depression symptom identification, in: Proceedings of the CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025.
[14] 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, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2019. Accessed: 2026-05-22.