UET-PsyWar@eRisk 2026: Weakly Supervised Multi-View Scoring for ADHD Sentence Ranking

Notebook for the eRisk Task 3 Lab at CLEF 2026

Duc-Luong Tran[1,†], Phuong-Anh Chu[1,†], Hoang-Dat Do[1,†], Tu-Phuong Mai[1], Duy-Cat Can[1,2,3] and Hoang-Quynh Le[1,*]

[1]VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam
[2]Platform of Bioinformatics, Lausanne University Hospital, Switzerland
[3]Faculty of Biology and Medicine, University of Lausanne, Switzerland

Abstract

In this work, we describe the UET-PsyWar team's approach to eRisk's 2026 Task 3: ADHD Symptom Sentence Ranking. Due to the nature of this task, which is characterized by the complete absence of sentence-level ADHD training labels, participants must overcome the severe lack of annotated clinical sentences mapping directly to the 18 ASRS-v1.1 symptoms. To address this, we propose a weakly supervised relevance filtering mechanism that leverages an external post-level dataset, utilizing attention pooling to extract sentence-level pseudo-labels and negative sampling from historical eRisk data. Following the filtering stage, candidate sentences are ranked using a multi-view scoring framework that combines an LLM-anchored semantic view, a cross-encoder diagnostic instruction view, and a rule-based lexical view. This hybrid approach aims to bridge the data scarcity gap and capture clinically meaningful evidence in the absence of official training data. Experimental results on the official test set under strict unanimity judgments show that our multi-view configuration achieved an Average Precision of 0.079 and an NDCG of 0.276. Evaluation of our individual and ensemble runs demonstrates that integrating explicit lexical markers with deep diagnostic reasoning outperforms isolated single-view retrieval methods.

Keywords

Attention-Deficit/Hyperactivity Disorder (ADHD), Sentence Ranking, Weak Supervision, Multi-view Scoring, Large Language Models

1. Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) screening traditionally relies on structured questionnaires such as the Adult ADHD Self-Report Scale (ASRS-v1.1) [1]. As social media becomes a viable medium for non-intrusive mental health detection [2, 3], eRisk 2026 Task 3 introduces the challenge of ADHD Symptom Sentence Ranking. Systems must evaluate user writings and retrieve up to 1,000 relevant sentences for each of the 18 ASRS-v1.1 symptoms.

Previous eRisk campaigns have seen a variety of methodological designs for identifying symptom-related sentences. Prior configurations have combined embedding clustering with multi-task transformer classifiers to predict symptom presence and continuous severity [4], or utilized two-stage filtering and re-ranking pipelines with lightweight sentence embeddings followed by larger retrieval models [5]. To combat severe class imbalances, other methodologies framed the ranking task as predicting a continuous numerical score for symptom severity, using ensembles of large language models (LLMs) to synthesize positive training examples for supervised fine-tuning [6].

The eRisk 2026 ADHD Symptom Sentence Ranking task presents three major challenges. First, there is no official training dataset for the 18 ADHD symptoms. Finding specific clinical markers in a massive, unannotated corpus is highly inefficient. To address this, a preliminary screening step is needed to discard irrelevant text and narrow down the search space. Second, standard workarounds for data

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
22021148@vnu.edu.vn (D. Tran); 23020324@vnu.edu.vn (P. Chu); 24020060@vnu.edu.vn (H. Do); 21020552@vnu.edu.vn (T. Mai); duy-cat.can@chuv.ch (D. Can); lhquynh@vnu.edu.vn (H. Le)

scarcity have severe limitations. Unsupervised similarity scoring is too noisy, while using LLMs to generate synthetic sentences produces artificial, unnatural text. To avoid this, systems should leverage structured clinical knowledge, such as diagnostic instructions and key vocabularies, rather than relying on synthetic data. Finally, there is an abstraction-level mismatch between formal clinical criteria and informal social media text. Diagnostic guidelines describe abstract behaviors, while social media posts describe specific, colloquial events. To bridge this gap, a semantic view is necessary to map everyday personal stories back to high-level clinical concepts. At the same time, a lexical view is required to ensure that explicit clinical keywords are not overlooked. Resolving this gap requires a hybrid scoring approach that can evaluate sentences from both a semantic and a lexical perspective.

In this working note, we describe our participation in Task 3 of the CLEF eRisk 2026 Lab [7, 8] under the team name *UET-PsyWar*. To address the lack of official training data and mitigate the abstraction-level mismatch, we propose a unified framework that couples a weakly supervised filtering step with a hybrid semantic-lexical scoring approach. By integrating this initial screening with semantic similarity, diagnostic instructions, and key vocabularies, our system is designed to identify and rank symptom-specific sentences from social media text.

2. Datasets

The official dataset provided for eRisk 2026 Task 3 consists of sentence-tagged documents derived from publicly available social media writings. These writings were specifically collected to contain ADHD-related expressions. Following the conventions of previous eRisk ranking tasks, the corpus assigns a unique identifier to each sentence, which is utilized for the final ranking submissions.

During the data ingestion phase, we applied a focused preprocessing strategy. We restricted our analysis strictly to the target sentence enclosed within the <TEXT> tag, deliberately excluding the surrounding conversational context provided by the <PRE> and <POST> tags. While this strict isolation risks losing broader conversational nuances, it was a necessary trade-off to computationally manage the massive corpus and minimize immediate noise during the initial ranking phase. Additionally, we applied basic text cleaning to remove URLs and hyperlinks before feeding the sentences into our proposed pipeline. The main statistics of the evaluation corpus are summarized in Table 1.

Table 1
Corpus statistics for eRisk 2026 Task 3: ADHD Symptom Sentence Ranking.

Metric	Value
Number of sentences	4,170,875
Average number of words per sentence	13.05
Number of ADHD symptoms	18

3. Proposed Method

To address the complete absence of annotated training data and the severe noise of social media text, we propose a two-phase pipeline as illustrated in Figure 1. In the first phase, a weakly supervised filter processes the raw corpus to discard irrelevant content and extract candidate sentences. In the second phase, a multi-view architecture is deployed to rank these extracted candidates. This ranking framework evaluates each candidate sentence across three complementary perspectives: an LLM-Anchored Semantic View, a Diagnostic Instruction View, and a Rule-based Lexical View. Specifically, the semantic view measures similarity against generated symptom seed sentences, the instruction view utilizes a cross-encoder to model deep logical entailment, and the lexical view bypasses semantic filtering to evaluate the exact presence of highly indicative clinical vocabulary. To initialize this multi-view evaluation, a large language model is leveraged beforehand to generate the required symptom-specific

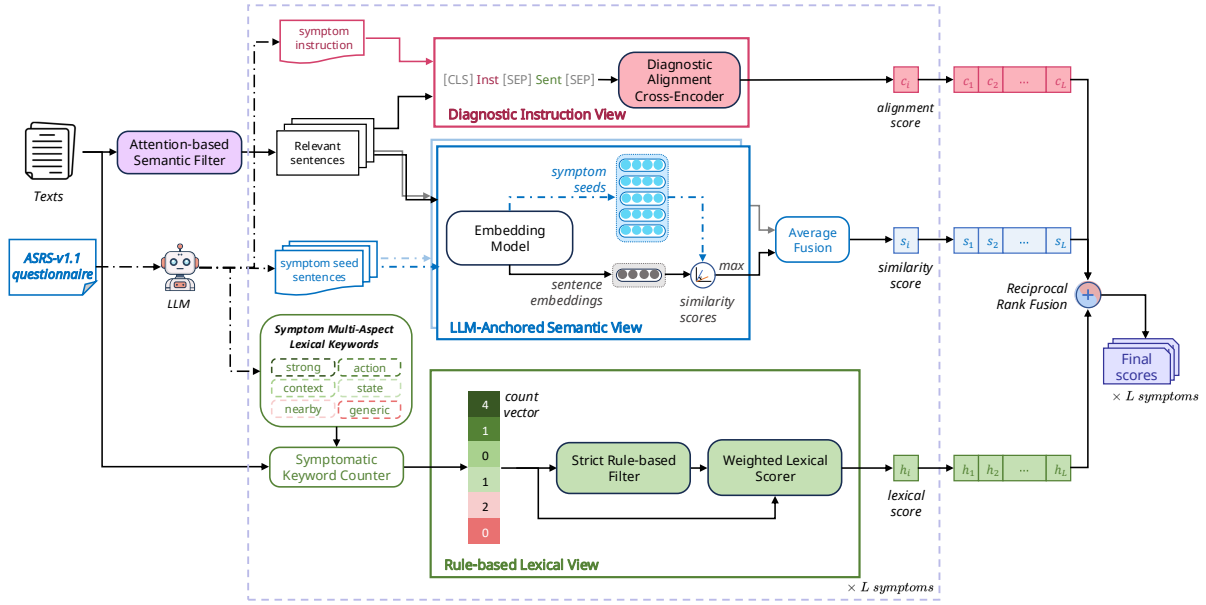

Figure 1: Architecture of the proposed multi-view scoring pipeline. The system utilizes an Attention-based Semantic Filter to extract relevant sentences from raw texts, guided by LLM-generated symptomatic priors. Candidates are independently scored through three parallel perspectives: a Diagnostic Instruction View (cross-encoder), an LLM-Anchored Semantic View (dense embeddings), and a Rule-based Lexical View (keyword tallying) before being aggregated via Reciprocal Rank Fusion for the final symptom-specific ranked lists.

diagnostic instructions, seed sentences, and multi-aspect keywords. Finally, the scores from these independent perspectives are aggregated to produce the final ranked list for each symptom.

3.1. Weakly Supervised Relevance Filtering

To extract candidate sentences from the unannotated corpus, we leverage an external dataset to implement a pseudo-labeling teacher-student architecture inspired by the Hierarchical Attention Network [9]. This model is bootstrapped on a binary corpus of approximately 11,300 Reddit posts (2,034 ADHD-positive) from the Mental Health Disorder Detection Dataset [10]. To project these post-level labels down to individual sentences, we employ an attention mechanism over a `roberta-large` encoder. For a post with N sentences, each sentence representation H_i is evaluated against a dynamic query vector u to compute a relevance score e_i:

$$e_i = u^T \tanh(WH_i + b) \tag{1}$$

where W and b are learnable parameters. These scores are normalized across the document using a Softmax function to obtain attention weights α_i:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N} \exp(e_j)} \tag{2}$$

The network optimizes the final document representation via attention pooling ($V = \sum_{i=1}^{N} \alpha_i H_i$), where N denotes the total number of sentences in the document. This weighted sum allows the network to emphasize sentences carrying strong diagnostic signals while suppressing irrelevant conversational noise. Sentences yielding an attention weight α_i above a dynamic density threshold γ that scales inversely with document length ($\gamma \propto 1/N$) are assigned a positive pseudo-label. This length-aware thresholding prevents the natural effect of Softmax attention weights spreading thin across many

sentences in longer documents, ensuring that valid clinical evidence is successfully extracted regardless of the post's overall verbosity, instead of a fixed threshold. Ultimately, this teacher model filtering process yields a refined set of positive pseudo-labeled sentences that represent high-probability ADHD clinical indicators.

To construct a balanced, discriminative training corpus for the sentence-level relevance filter, we augment these pseudo-labeled sentences with negative samples drawn from historical eRisk depression corpora. Using this weakly labeled dataset, we train a standalone Student model to act as the primary inference filter. This model relies on a `roberta-large` backbone, passing the `[CLS]` token representation through a dropout layer ($p = 0.5$) and a linear classification head. The network is optimized using `BCEWithLogitsLoss`, incorporating a dynamic positive class weight to mitigate the inherent imbalance of symptomatic versus benign sentences. During the final inference phase on the raw eRisk 2026 test data, sentences that yield a sigmoid probability higher than 0.5 are classified as relevant and forwarded to the multi-view scorers.

3.2. Multi-view Symptom-specific Scoring

Once the candidate sentences have passed through the weakly supervised relevance filter, they are evaluated to determine their specific alignment with the 18 ASRS-v1.1 symptoms. Because clinical symptoms can manifest in highly varied and implicit ways in social media text, relying on a single scoring mechanism is often insufficient. Therefore, we deploy a multi-view scoring architecture that evaluates each sentence from semantic, diagnostic, and lexical perspectives.

3.2.1. LLM-anchored Semantic View

Clinical questionnaire language often differs significantly from colloquial social media text. To bridge this gap, we utilize ChatGPT 5.4 to generate diverse, first-person seed sentences that mirror the self-reported nature of personal mental health disclosure, acting as linguistic anchors for each symptom. Formally, we define a target symptom l anchored by K generated seeds $\mathcal{Q}_l = \{q_{l,1}, \ldots, q_{l,K}\}$, alongside a filtered candidate sentence $x \in \mathcal{C}$. To project these texts into a shared latent space, we deploy two parallel dense encoders (`nomic-embed-text` and `ModernBERT`) to obtain diverse semantic representations and reduce single-model bias. We then compute the cosine similarity between the dense vector of the candidate sentence and the vectors of the generated seeds; this metric quantifies their semantic proximity, allowing us to capture informal paraphrases that share underlying clinical meaning despite lacking exact vocabulary overlap.

First, for a given embedding model M, we compute the cosine similarity between the candidate sentence x and each of the K seeds, resulting in K similarity scores. We then apply max-pooling to select the single highest score, as a relevant candidate only needs to strongly align with *one* specific symptom manifestation. This yields one maximum score per model, with E_M representing the embedding function of model M:

$$s_{l,M}(x) = \max_{k=1}^{K} \Big(\text{sim}(E_M(x), E_M(q_{l,k})) \Big) \tag{3}$$

Applying this pooling process to both encoders yields two separate similarity scores: `nomic-embed-text` ($s_{l,\text{nomic}}(x)$) and `ModernBERT` ($s_{l,\text{mbert}}(x)$). Finally, an Average Fusion step calculates the mean of these two scores to produce a robust final semantic score, $s_l(x)$:

$$s_l(x) = \frac{s_{l,\text{nomic}}(x) + s_{l,\text{mbert}}(x)}{2}. \tag{4}$$

3.2.2. Diagnostic Instruction View

While the semantic view efficiently captures broad topical similarities, bi-encoder architectures often compress complex clinical criteria into single vectors, potentially losing fine-grained token-level relationships. To address this, we incorporate a Diagnostic Instruction View powered by the Alibaba GTE (General Text Embedding) cross-encoder [11]. Unlike bi-encoders, cross-encoders process the

query and candidate text simultaneously, applying full self-attention across the entire sequence to capture deep, token-level interactions. We utilize the pre-trained GTE model because of its robust performance in text retrieval and re-ranking, which is achieved through extensive weakly supervised pre-training on massive text-pair datasets followed by supervised contrastive fine-tuning. This allows our pipeline to leverage its powerful alignment capabilities to evaluate diagnostic relevance without requiring domain-specific retraining.

In practice, this view evaluates a concise, LLM-generated diagnostic instruction $\mathcal{I}_l$, along with the filtered candidate sentence $x \in \mathcal{C}$. The cross-encoder processes the concatenated sequence to output an alignment score, $c_l(x)$:

$$c_l(x) = \text{GTE}([\text{CLS}]\ \mathcal{I}_l\ [\text{SEP}]\ x\ [\text{SEP}]) \tag{5}$$

By modeling their token-level dependencies, GTE outputs this alignment score as a highly discriminative signal of how the text satisfies the clinical criteria.

3.2.3. Rule-based Lexical View

While dense embeddings and cross-encoders excel at capturing implicit semantics, they occasionally dilute explicit, highly indicative clinical vocabulary. To guarantee these exact signals are preserved, we introduce a Rule-based Lexical View. Bypassing the initial semantic filter, this view operates directly on the raw text to evaluate the exact presence of LLM-generated, multi-aspect lexical keywords.

For each candidate sentence x, a keyword counter tallies the occurrences of terms across six predefined aspects, generating a count vector: $[\text{strong}, \text{action}, \text{context}, \text{state}, \text{nearby}, \text{generic}]$. This vector is first evaluated by a Strict Rule-based Filter to eliminate obvious false positives. Instead of complex nested conditions, the filter applies straightforward boolean logic: it rejects sentences dominated by penalty terms (e.g., generic ≥ 2 or nearby ≥ 2 when strong $= 0$), and accepts sentences that either contain at least one *strong* indicator (strong ≥ 1) or exhibit specific co-occurring aspects (such as combinations of *state*, *action*, and *context* terms).

For the remaining sentences, the final lexical score $h_l(x)$ is computed using a clinically inspired heuristic weighting scheme. The coefficients are empirically calibrated to reflect diagnostic specificity: definitive symptom expressions receive aggressive rewards, whereas ambiguous or contextually negated terms incur heavy penalties. Furthermore, the contribution of *state* terms is capped to prevent score inflation from repetitive phrasing:

$$h_l(x) = 4.0\times\text{strong}+1.6\times\text{action}+1.2\times\text{context}+0.8\times\min(\text{state}, 2)-1.5\times\text{nearby}-2.0\times\text{generic} \tag{6}$$

3.3. Score Normalization & Submitted Configurations

Because the proposed views output scores on entirely different scales, ranging from bounded cosine similarities to unbounded cross-encoder logits and lexical heuristic counts, direct aggregation of raw scores is not feasible. To resolve this, we convert the raw scores from each view into normalization strategies based on the chosen fusion method. For linear combination strategies, we apply min-max normalization to the raw scores of each view into a uniform $[0, 1]$ range. For fusion strategies based on Reciprocal Rank Fusion (RRF) [12], we directly utilize the raw integer ranks.

To evaluate the independent and synergistic contributions of our multi-view framework, our team, *UET-PsyWar*, submitted five distinct configurations to the eRisk evaluation server:

To formalize our aggregation strategies, let $\tilde{r}_{\text{sim}}$, $\tilde{r}_{\text{ce}}$, $\tilde{r}_{\text{key}}$, and $\tilde{r}_{\text{sem}}$ denote the Min-Max normalized scores (scaled to $[0, 1]$) for the semantic similarity, cross-encoder, lexical, and initial semantic filter scores, respectively. Similarly, let r_{sim}, r_{ce}, r_{key}, and r_{sem} denote their corresponding raw integer ranks. We submitted the following five configurations:

- **Run1 (sim only):** Evaluates the isolated performance of the LLM-anchored Semantic View.

$$\text{Score} = \tilde{r}_{\text{sim}}$$

- **Run2 (CE only):** Evaluates the isolated performance of the Diagnostic Instruction View powered by the cross-encoder.

$$\text{Score} = \tilde{r}_{\text{ce}}$$

- **Run3 (balanced hybrid):** Employs a weighted sum of normalized ranks from all components, heavily favoring the cross-encoder's diagnostic alignment.

$$\text{Score} = 0.45 \cdot \tilde{r}_{\text{ce}} + 0.45 \cdot \tilde{r}_{\text{sim}} + 0.10 \cdot \tilde{r}_{\text{sem}}$$

- **Run4 (hybrid key bonus):** Employs a weighted sum of all normalized ranks, shifting the dominant weight to the semantic similarity view and lexical features to test broad semantic matching against strict diagnostic criteria.

$$\text{Score} = 0.40 \cdot \tilde{r}_{\text{ce}} + 0.40 \cdot \tilde{r}_{\text{sim}} + 0.10 \cdot \tilde{r}_{\text{sem}} + 0.10 \cdot \tilde{r}_{\text{key}}$$

- **Run5 (RRF robust):** Bypasses linear rank normalization to apply a Weighted Reciprocal Rank Fusion using the raw ranks. With a smoothing constant of $k = 60$, following standard RRF practices, this run integrates all views while heavily prioritizing the cross-encoder and semantic similarity components.

$$\text{Score} = \frac{0.45}{60 + r_{\text{ce}}} + \frac{0.45}{60 + r_{\text{sim}}} + \frac{0.10}{60 + r_{\text{sem}}}$$

4. Evaluation Results

In this section, we present the evaluation results of our proposed framework, which were computed by the organizers on the official eRisk 2026 test dataset. Our submitted configurations are assessed using standard information retrieval metrics: Average Precision (AP), R-Precision (R-PREC), Precision at 10 (P@10), and Normalized Discounted Cumulative Gain at 1000 (NDCG@1000) [13, 14].

4.1. Comparison with Participating Teams

To contextualize our performance, Table 2 presents the results of our best configuration alongside the top-performing runs from other participating teams.

Table 2
Evaluation results for Task 3 of the best runs from top teams under **unanimity** judgments

Team	Run	AP	R-PREC	P@10	NDCG@1000
NeuroRankUB	final ADHD ranking	**0.207**	**0.274**	**0.367**	**0.488**
erisk-cedri	EnsembleV3	0.138	0.193	0.261	0.387
INSA-Lyon	Ensemble	0.129	0.160	0.244	0.363
DS-GT	asrs run5	0.080	0.121	0.217	0.188
UET-PsyWar	run4 hybrid key bonus	0.079	0.122	0.156	0.276
AI-MH-Z	ST equal weights	0.078	0.108	0.167	0.291

Shaded row indicates our submission. Best result of each metric is highlighted in bold.

As shown in Table 2, our multi-view framework places solidly in the middle tier of the participating systems. The leading team, NeuroRankUB, achieved an AP of 0.207, establishing a significant gap between the top-tier systems and the rest of the participants. We hypothesize that the primary reason for this performance gap stems from a bottleneck in our weakly supervised relevance filter. In our architecture, all candidate sentences must pass through this initial filter (which applies a strict probability threshold of 0.5) before reaching the multi-view scorers. If this filter was overly aggressive, it likely discarded subtle, implicit, but clinically valid ADHD sentences early in the pipeline. Consequently, the downstream cross-encoder and semantic scorers never had the opportunity to rank these sentences, inherently capping our maximum achievable recall and Average Precision.

4.2. Internal Run Comparison

To evaluate the independent and synergistic contributions of our proposed views. Table 3 details the performance of our five submitted configurations. Comparison within our internal runs also provides valuable ablation insights into the components of our multi-view framework.

Table 3
Evaluation results for Task 3 of our configurations under **unanimity** judgments

Team	Run	AP	R-PREC	P@10	NDCG@1000
UET-PsyWar	run1 sim only	0.020	0.044	0.050	0.149
UET-PsyWar	run2 CE only	0.058	0.099	0.161	0.242
UET-PsyWar	run3 balanced hybrid	0.071	**0.128**	0.167	0.268
UET-PsyWar	run4 hybrid key bonus*	**0.079**	0.122	0.156	**0.276**
UET-PsyWar	run5 RRF robust	0.072	0.113	**0.200**	0.262

Best configuration is highlighted by *. Best result of each metric is highlighted in bold.

Cross-Encoder vs. Semantic Similarity: Across both evaluation settings, the isolated Diagnostic Instruction View (run2 CE only) significantly outperformed the LLM-anchored Semantic View (run1 sim only) across all metrics. For instance, under unanimity judgments, run2 CE only achieved an AP of 0.058 compared to run1 sim only's 0.020. This highlights the architectural advantages of the cross-encoder. While the semantic similarity view compresses candidate sentences and symptom seeds into single, isolated vectors, the Alibaba GTE cross-encoder performs full self-attention across the concatenated instruction-sentence pair. Its pre-training on massive question-answer datasets equips it with superior token-level reasoning capabilities, allowing it to better discern whether a sentence genuinely satisfies complex diagnostic criteria.

Ensemble Performance and Lexical Importance: Our ensemble configurations (run3 balanced hybrid, run4 hybrid key bonus, and run5 RRF robust) consistently outperformed the single-view runs, validating the hypothesis that combining diagnostic, semantic, and lexical signals yields a more robust ranking. Under the strict unanimity judgments, run4 hybrid key bonus emerged as our best overall configuration. Because Run 4 (hybrid key bonus) assigns higher weights to the semantic similarity and lexical keyword features, this suggests that when human experts unanimously agree on a sentence's relevance, the sentence is highly likely to contain explicit, obvious lexical markers of the symptom, features that Run 4 (hybrid key bonus) is specifically optimized to detect. Furthermore, while run5 RRF robust achieved the highest early precision (P@10: 0.200) under unanimity, the linear combination in run4 hybrid key bonus provided the best overall ranking quality (NDCG@1000: 0.276).

5. Conclusion

In this paper, the UET-PsyWar team addressed the lack of sentence-level training data in eRisk 2026 Task 3 by proposing a multi-view ranking framework. Our pipeline utilized a weakly supervised relevance filter to extract candidate sentences, which were subsequently evaluated through LLM-anchored semantic, cross-encoder diagnostic, and rule-based lexical views. Experimental results confirmed that our ensemble configurations consistently outperformed isolated single-view approaches, with our lexical-semantic combination (Run 4) specifically revealing a strong correlation between unanimous human agreement and explicit lexical markers. This performance ultimately demonstrates our system's solid competitive viability with opportunities for optimization in clinical sentence retrieval.

Despite the internal robustness of our multi-view scorers, our system placed in the middle tier overall, revealing critical pipeline flaws. Primarily, relying on post-level attention weights for sentence-level pseudo-labels introduced noise. Applying a strict probability threshold on this filter likely acted as a bottleneck, discarding subtle ADHD expressions and severely capping maximum recall. Additionally, strictly isolating the target sentence while discarding the surrounding conversational context removed crucial interpretive clues. Finally, our semantic and lexical views relied heavily on LLM-generated

seeds and keywords, which occasionally produced overly clinical phrasing that failed to align with colloquial social media text. Future work will address these limitations by implementing soft-filtering mechanisms, integrating local conversational context, and refining LLM prompting strategies to better capture natural psychiatric linguistic variations.

Acknowledgments

This work was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under the project "Research and Development of a Personalized Machine Learning System for Early Alzheimer's Disease Diagnosis from Adaptive Multimodal Biomarkers" (Grant No. 102.05-2025.54).

Declaration on Generative AI

During the preparation of this report, Grammarly and ChatGPT were used solely for grammar correction and improving the readability of the manuscript.

References

[1] R. C. Kessler, L. Adler, M. Ames, O. Demler, S. Faraone, E. Hiripi, M. J. Howes, R. Jin, K. Secnik, T. Spencer, T. B. Ustun, Walters, E. E, 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.

[2] S. Chancellor, M. De Choudhury, Methods in predictive techniques for mental health status on social media: a critical review, npj Digital Medicine 3 (2020) 43. doi:10.1038/s41746-020-0233-7.

[3] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J. Dobson, R. Dutta, Characterisation of mental health conditions in social media using Informed Deep Learning, Scientific reports 7 (2017) 45141. doi:10.1038/srep45141.

[4] T.-P. Mai, M.-H. H. Le, D.-L. Tran, D.-C. Can, H.-Q. Le, UET@eRisk2025: Severity Estimation for Depression Symptoms Searching and Early Risk Detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038, CEUR-WS.org, 2025, pp. 1550–1561.

[5] N. M. Son, D. V. Thin, SonUIT eRisk2025: Enhanced Depression Detection on Social Media via Filtering and Re-Ranking, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038, CEUR-WS.org, 2025, pp. 1583–1589.

[6] D. A. P. Nunes, E. Ribeiro, INESC-ID @ eRisk 2025: Exploring Fine-Tuned, Similarity-Based, and Prompt-Based Approaches to Depression Symptom Identification, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038, CEUR-WS.org, 2025, pp. 1474–1485.

[7] 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.

[8] 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.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical Attention Networks for Document Classification, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1480–1489. URL: https://aclanthology.org/N16-1174/. doi:10.18653/v1/N16-1174.

[10] N. S. Rupol, Mental Health Disorder Detection Dataset, 2023. URL: https://www.kaggle.com/datasets/nazmussakibrupol/mental-health-disorder-detection-dataset.

[11] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023. URL: https://arxiv.org/abs/2308.03281. arXiv:2308.03281.

[12] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[13] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, volume 35, Cambridge University Press, 2008. doi:10.1162/coli.2009.35.2.307.

[14] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: https://doi.org/10.1145/582415.582418. doi:10.1145/582415.582418.