VANGUARD at eRisk 2026 Task 3: ADHD Symptom Sentence Ranking

Notebook for the eRisk Lab at CLEF 2026

Anamaria Vișan[1], Ciprian-Octavian Truică[1,2], Adrian Paschke[3] and Elena-Simona Apostol[1,*]

[1]National University of Science and Technology Politehnica Bucharest, Splaiul Independenței 313, București 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania
[3]Fraunhofer Institute for Open Communication Systems, Berlin, 10589, Germany

Abstract

In this paper, we describe team VANGUARD at eRisk 2026 Task 3, the first edition of ADHD Symptom Sentence Ranking. Participants rank approximately 4.17 million social-media sentences by relevance to each of 18 ASRS-v1.1 symptoms under a polarity-agnostic criterion and without in-domain training labels. We implement a zero-shot hybrid IR pipeline with BM25 multi-query retrieval, dense bi-encoder similarity, MS MARCO-style cross-encoder reranking, optional Ollama/LangChain LLM rescoring, and reciprocal rank fusion (RRF). We submitted four TREC runs; leaderboard figures in the lab overview [1, 2] were still pending at publication time. On released majority-vote qrels, our best nDCG@1000 is 0.113 with the cross-encoder run.

Keywords

ADHD, information retrieval, social media, ASRS, eRisk, hybrid retrieval, cross-encoder

1. Introduction and task description

1.1. Motivation and task overview

Task 3 (2026), ADHD Symptom Sentence Ranking, is an eRisk track that targets sentence-level retrieval for the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1). Participants rank candidate sentences by estimated relevance to each symptom separately, rather than assigning a single ADHD vs. not label. By extending symptom-oriented retrieval beyond depression to ADHD, the task supports interpretable, symptom-aware search and cross-condition comparison at sentence granularity [1, 2].

1.2. Relevance criterion

A sentence is relevant when it conveys information about the user's state with respect to the target ADHD symptom, irrespective of polarity or stance. For example, both inability and ability to focus count for attention symptoms because both describe lived experience of the symptom domain. The design encourages models to surface clinically meaningful evidence, not only keywords or negative sentiment. This is an IR task rather than a diagnostic classifier, and the resulting rankings are research artifacts evaluated against pooled expert judgments.

1.3. Data release

Organizers provide a sentence-tagged corpus (∼4.17M sentences in 4,521 topic files) derived from publicly available social-media writings collected to contain ADHD-related expressions, using the

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
anamaria.visan@stud.acs.upb.ro (A. Vișan); ciprian.truica@upb.ro (C. Truică); adrian.paschke@fokus.fraunhofer.de (A. Paschke); 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)
0009-0002-7229-9649 (A. Vișan); 0000-0001-7292-4462 (C. Truică); 0000-0003-3156-9040 (A. Paschke); 0000-0001-6397-4951 (E. Apostol)

same broad TREC-style conventions as recent eRisk ranking tasks. As the first edition of this ADHD ranking task, no annotated in-domain training data are released. Organizers provide only test inputs and the ASRS-v1.1 symptom definitions, so systems typically rely on zero-shot or externally pretrained retrievers.

1.4. Submission stage

After data release, teams produce up to five independent runs (five systems) and upload TREC-formatted rankings to the organizer FTP. For each symptom $s \in \{1, \ldots, 18\}$, a run contains up to 1,000 lines ordered by decreasing estimated relevance. Each line encodes the symptom number and the literal field `Q0`, followed by the sentence identifier, rank position, score, and system name. For example, symptom 1 appears at ranks 1 through 1,000 with decreasing scores, and the same structure repeats for all 18 symptoms. Only rank order within a run is used for evaluation, and score scales may differ across systems.

1.5. Evaluation stage

When submissions close, organizers build relevance judgments by top-k pooling of high-ranked sentences across participant runs, followed by expert annotation. Resulting qrels support standard IR metrics, including MAP, nDCG@1000, and P@10. Lab outcomes are published in CEUR-WS working notes.

1.6. Core difficulties and scale

Naïve scoring of every sentence for 18 symptoms is infeasible at corpus scale, so practical systems use multi-stage retrieval with a sparse or dense first stage followed by cross-encoder or LLM reranking on shortlists. Informal language, symptom overlap, and polarity-agnostic relevance increase the need for semantic matching beyond keyword overlap.

1.7. Our participation (VANGUARD)

We address Task 3 with a hybrid zero-shot pipeline that uses BM25, a dense bi-encoder, cross-encoder reranking, optional local LLM rescoring, and reciprocal rank fusion.

The notebook computes five rankers (Runs 1 to 5), but only four were archived and officially submitted: Run 1 (BM25), Run 3 (cross-encoder), Run 4 (LLM rerank, partial), and Run 5 (RRF fusion). Run 2, the dense bi-encoder/semantic ranker, was computed but not included in the submission archive. Run 5's reciprocal rank fusion combines only two ranked lists, from BM25 and the cross-encoder, instead of the three that a dense-semantic list would have allowed (Section 3).

Section 4 reports submission statistics, evaluation on released qrels, and organizer participation counts reported in the lab overview [1].

2. State of the art

Our pipeline design is grounded in a structured synthesis of 29 papers on ADHD language, social-media mental-health NLP, neurodevelopmental technology reviews, and information retrieval. The corpus spans user-level classification studies, systematic reviews, hybrid and explainable models, and large-scale benchmarks. In the following, we distill the themes most relevant to Task 3. These include eighteen ASRS-v1.1 symptom queries, polarity-agnostic sentence relevance, and zero-shot ranking on corpus scale (approximately 4.17 million sentences) under MAP, nDCG@1000, and P@10.

2.1. ADHD and mental-health language on social media

Guntuku et al. show that adult ADHD communities on Twitter differ from matched controls in LIWC and open-vocabulary features related to self-focus, affect, temporal orientation, and themes such as emotional dysregulation and exhaustion (user-level AUC $\sim$0.84) [3]. Chen et al. combine content and interaction metadata on Twitter, arguing that behavior patterns complement text alone [4]. Qualitative NLP work extracts themes of lived experience from social posts. These themes show partial overlap with DSM-5 and stronger alignment with clinical manuals, supporting symptom-oriented rather than single-label screening, as also argued in ADHD-centric detection work [5]. Task 3 operationalizes this at sentence granularity using ASRS-v1.1 symptom definitions as queries, with relevance defined irrespective of stance (ability and inability to focus both count).

2.2. Disorder classification vs. symptom-level retrieval

Most social media studies treat mental health detection as a multiclass classification problem. RoBERTa on Reddit separates depression, anxiety, bipolar, ADHD, PTSD, and control classes [6]; hybrid models combine psycholinguistic features with transformers for six conditions [7]. Such setups often report high F1 on self-disclosed labels. They nevertheless conflate disorder identity with symptom expression.

Recent surveys describe growing use of LLMs, RAG, and agentic pipelines, together with concerns about trustworthiness and hallucination [8]. LLM and classical ensembles on narrative transcripts can improve ADHD detection recall, but remain binary and clinically unvalidated [9]. Task 3 differs: it is an IR task with eighteen independent ranked lists, expert qrels from pooled submissions, and no in-domain training labels.

2.3. Systematic reviews: social data, imaging, and technology

Rashti Mohammad et al. review ML on social media for NDDs. Most studies cover ASD and ADHD on Reddit, Twitter, YouTube, and Facebook; reported F1 ranges from 0.48 to 0.89, but samples are small, and ethics concerns recur [10], while class imbalance poses a major limitation [11].

Imaging reviews offer background on validation and interpretability, yet they are largely orthogonal to our text-only setting [12]. Ribas et al. synthesize 221 studies on technology for NDD diagnosis and treatment and note both promise and high risk of bias, with limited ADHD-specific NLP guidance [13]. Large Reddit benchmarks improve cleaning and baselines but still rely on disorder-level labels [6]. Task 3 releases only test sentences and ASR text; we therefore use pretrained retrievers and hand-built symptom queries.

2.4. Hybrid, sentiment, and explainable approaches

Hybrid pipelines that combine engineered features with BiLSTM or transformers remain competitive [7]. Kerz et al. compare interpretable feature models with mental-health transformers using LIME/AGRAD and include ADHD among the benchmark conditions [14]. Wiechmann et al. report on ADHD-centric Reddit/Twitter classifiers and cross-platform generalization [5].

Sentiment alone is a poor proxy for ADHD. Under Task 3's polarity-agnostic criterion, positive and negative mentions of a symptom domain are both relevant. We therefore report ranking effectiveness under IR metrics and keep BM25 and fusion stages auditable.

2.5. Information retrieval and multi-stage ranking

Scoring every sentence against 18 symptoms is infeasible at this scale. Multi-stage retrieval is the practical alternative.

A typical stack starts with sparse lexical matching (BM25 or TF-IDF), adds dense bi-encoders for semantic similarity [15], reranks shortlists with cross-encoders trained on MS MARCO, and may fuse ranked lists without calibrating scores across systems [16]. BEIR-style benchmarks illustrate zero-shot

dense retrieval under domain shift [17]. Local LLM reranking (Ollama or LangChain) follows similar ideas but adds latency. eRisk Task 3 adopts TREC-style submission, top-k pooling, and expert qrels, with nDCG@1000 as the official cut-off [1, 2].

2.6. Design implications for Task 3

The synthesis supports our run portfolio. Run 1 applies lexical multi-query BM25 as a sparse first stage. Run 2 encodes the full corpus with a dense bi-encoder to obtain semantic similarity scores. Run 3 reranks BM25 candidates with a cross-encoder trained on MS MARCO-style passage ranking. Run 4 optionally rescores shortlists with a local LLM under polarity-agnostic instructions. Run 5 fuses the ranked lists with reciprocal rank fusion and assigns higher weight to cross-encoder rankings. We instantiate symptom queries from ASRS questions, short labels, and expanded keywords, and our multi-query score fusion mirrors hybrid retrieval practice in the literature. Because of compute limits, Run 4 exports only 40 lines per symptom, whereas Runs 1, 3, and 5 provide complete 1,000-line rankings.

2.7. Ethics and limitations

Across the 29-paper synthesis, authors stress that social-media models are not diagnostic tools. Self-report and subreddit proxies introduce label noise, platform- and language-specific bias limits generalization, and privacy and stigma constrain data collection [10, 3]. Ranked sentences in Task 3 are challenge artifacts for pooled evaluation and must not be treated as evidence of ADHD in individuals.

3. Proposed solution

Team VANGUARD implemented Task 3 in a single notebook pipeline that loads the organizer TREC sentence corpus, builds an ASRS-v1.1 query profile per symptom, runs up to five independent rankers, and exports VANGUARD_run{N}.trec files for FTP submission.

3.1. Architecture and corpus loading

The TRECDataLoader parses `output_trec_files/s_*.trec` documents with <DOCNO>, <TEXT>, and optional <PRE>/<POST> fields. It keeps sentences between 10 and 500 characters and deduplicates document identifiers. After filtering and deduplication, our run loaded 4,170,875 sentences from 4,521 TREC topic files (median length 58 characters, mean 71.3, maximum 500).

The ASRS symptom table contains 18 entries. Each entry stores the official ASRS question together with a short label, a domain classification (inattention, hyperactivity, or impulsivity), expanded keywords, and a hand-built query string for lexical retrieval. We frame relevance as polarity-agnostic, meaning that any mention of the symptom domain counts whether the user reports having or lacking the trait.

Our configuration exports up to 1,000 lines per symptom, retains 5,000 BM25 candidates for reranking, and cross-encoder reranks the top 1,000 BM25 hits per symptom. The embedding model is `sentence-transformers/all-mpnet-base-v2` (batch 256, normalized vectors), and the cross-encoder is `ms-marco-MiniLM-L-6-v2` (batch 256, maximum length 256).

3.2. Run 1: BM25 (sparse)

The BM25Ranker tokenizes the corpus with BM25Okapi. For each symptom, we score three queries and fuse them as $0.4\times$ query + $0.35\times$ keywords + $0.25\times$ the lowercased ASRS question. We retain up to 5,000 candidates for downstream stages and export the top 1,000 under the system name VANGUARD_BM25.

3.3. Run 2: Dense bi-encoder

The SemanticRanker encodes all sentences once and ranks them by cosine similarity to the mean of four query embeddings derived from the ASRS question, the short label, the expanded query, and the

Table 1
Submitted TREC runs (VANGUARD).

Run	System	Lines	/ symp.	Score range (pooled)
1	BM25	18,000	1,000	7.16 to 27.14
2	MPNet semantic	n/a	n/a	(not in archive)
3	Cross-encoder	18,000	1,000	−11.45 to 2.86
4	LLM (Ollama)	720	40	2.08 to 8.60
5	RRF ensemble	18,000	1,000	0.002 to 0.062

template `ADHD symptom: {short}`. The notebook saves this run as `VANGUARD_Semantic`, but it is not included in our archived submission folder (see Section 4).

3.4. Run 3: Cross-encoder reranking

The CrossEncoderReranker retrieves the top 1,000 BM25 candidates per symptom, scores each (question, sentence) pair with the MS MARCO cross-encoder, and exports the top 1,000 lines as `VANGUARD_CrossEncoder`.

3.5. Run 4: LLM rerank (Ollama / LangChain)

The LLMReranker takes the top 40 BM25 sentences per symptom and prompts Ollama (`llama3.2`) through LangChain's OllamaLLM for a relevance score from 0 to 10 under a polarity-agnostic instruction. The final score combines $0.3\times$ normalized BM25 (scaled to 10) with $0.7\times$ the LLM score. The export contains 720 lines (18×40) under `VANGUARD_LLM`, which is below the task maximum of 1,000 per symptom.

3.6. Run 5: Reciprocal rank fusion

The ReciprocalRankFusion module (k=60) merges ranked lists using $\mathrm{RRF}(d) = \sum_r w_r/(k+\mathrm{rank}_r(d))$. Our submitted fusion combines BM25 and cross-encoder lists with weights 1.0 and 1.5. Dense semantic fusion is implemented but was disabled in the final export, and LLM lists receive weight 1.3 when that stage completes. We export the fused ranking as `VANGUARD_RRFEnsemble`.

3.7. TREC export and execution flow

The TRECOutputFormatter writes tab-separated lines in the form the standard TREC line format (symptom, `Q0`, document id, rank, score, system name), up to 1,000 entries per symptom for team VANGUARD. The notebook executes stages sequentially. It first builds the BM25 index and exports Run 1, then encodes the corpus with MPNet for Run 2, applies cross-encoder reranking for Run 3, optionally installs and serves Ollama for Run 4, and finally computes RRF for Run 5. Run 4 requires a local Ollama service for `llama3.2`, and Colab helper cells install and start the daemon when needed.

4. Results

Table 1 summarizes TREC files in our submission archive (four of five notebook runs exported).

Each of the 18 symptom topics contributes up to 1,000 ranked lines (40 for Run 4). Score scales are not comparable across runs, and the evaluation uses order only.

4.1. Run-level behaviour

Run 1 (BM25) provides a lexical baseline with overlap on symptom phrases such as "fidget" and "interrupt". Run 3 (cross-encoder) reranks BM25 shortlists and yields the highest MAP and nDCG@1000

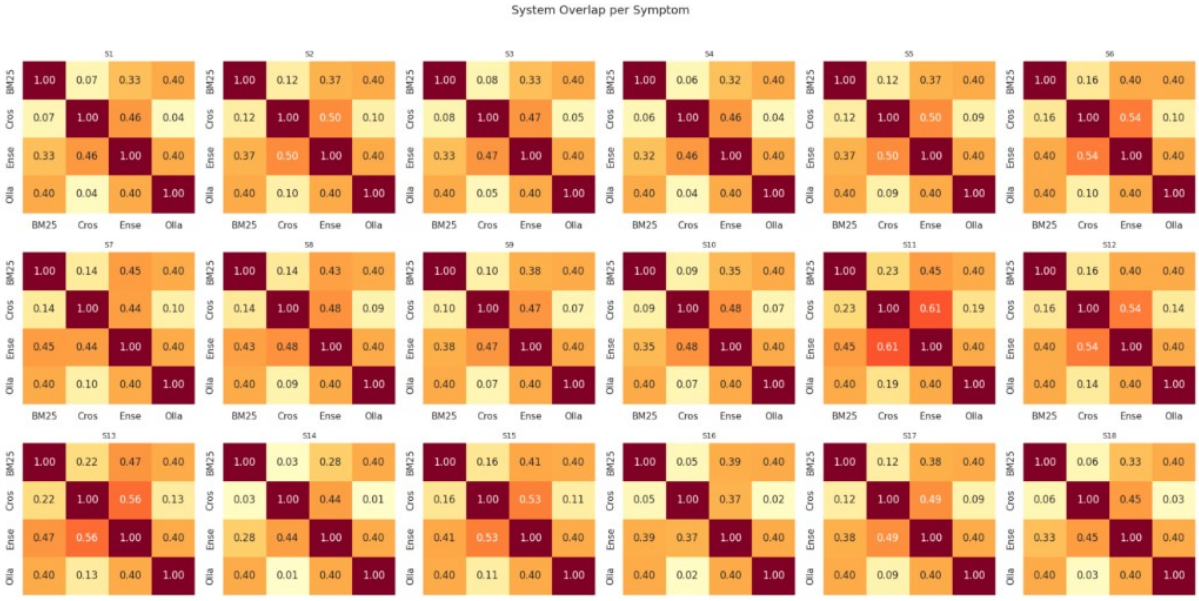

Figure 1: Pairwise top-100 ranking overlap (Jaccard similarity) per ASRS symptom (S1 to S18) for BM25, cross-encoder (Cros), RRF ensemble (Ense), and Ollama (Olla).

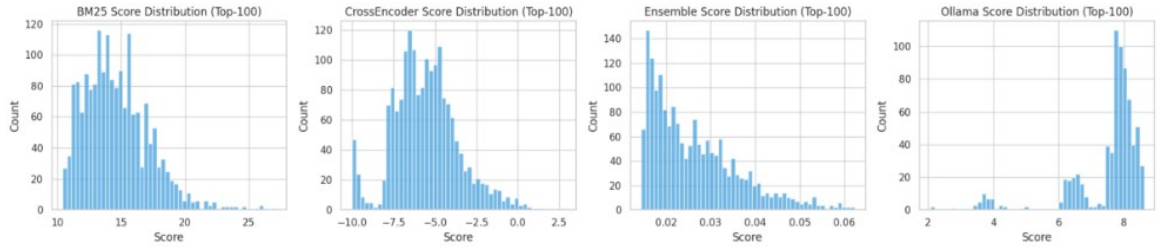

Figure 2: Histogram of top-100 scores pooled over all 18 symptoms for BM25, cross-encoder, RRF ensemble, and Ollama.

in Table 2. Run 5 (RRF) fuses BM25 and cross-encoder lists while omitting the dense semantic export from Run 2. Run 4 scores only 40 BM25 candidates per symptom via Ollama, which is useful for ablation but under-submits relative to the 1,000-line task definition.

4.2. Ranking overlap across systems

Figure 1 compares the top-100 ranked sentences for each symptom across BM25, cross-encoder, RRF ensemble, and Ollama runs. Overlap is measured as Jaccard similarity between document-ID sets ($|A \cap B|/|A \cup B|$). BM25 and Ollama agree on roughly 40% of top-100 documents on most symptoms, consistent with the LLM stage reranking a BM25 shortlist. Cross-encoder rankings overlap less with BM25 (often below 0.15), indicating that semantic reranking retrieves different sentences. The ensemble shows intermediate overlap with both BM25 and the cross-encoder (typically 0.30 to 0.55), reflecting the weighted fusion of both lists. Symptom-level variation is visible: for example, S11 shows higher cross-encoder/ensemble agreement (0.61), whereas S14 shows very low cross-encoder overlap with BM25 and Ollama (0.03 and 0.01).

Figure 2 pools the top-100 scores across all 18 symptoms for each submitted system. BM25 scores are positive and roughly bell-shaped (about 10 to 27). Cross-encoder logits are negative and concentrated around -6, with a small tail near -10. RRF fusion produces much smaller values (about 0.015 to 0.065) with most mass at the low end, as expected from reciprocal-rank contributions. Ollama relevance scores (from 0 to 10) cluster around 7.5 to 8.5, with smaller modes near 4 and 6.5. These distributions are not comparable across runs; evaluation uses rank order only.

Table 2
Task 3 metrics vs. majority-vote qrels (macro-average over 18 symptoms).

Run	MAP	nDCG@1000	P@10
1 BM25	0.014	0.088	0.056
3 Cross-encoder	0.028	0.113	0.156
4 LLM (40 docs)	0.014	0.052	0.083
5 RRF ensemble	0.023	0.107	0.111

4.3. Evaluation against released qrels

Organizers released majority-vote and unanimity relevance judgments (qrels) for pooled assessment. We computed IR metrics on our TREC submissions (Table 2); the official cut-off is nDCG@1000.

Cross-encoder reranking (Run 3) achieves the highest MAP and nDCG@1000, whereas BM25 alone is weakest. Run 4 (40 candidates/symptom) underperforms full 1,000-line runs, because nDCG@1000 credits relevant documents at ranks up to 1,000, capping the export at 40 lines per symptom mechanically bounds recall, so Run 4's lower score partly reflects list length rather than ranking quality alone.

RRF (Run 5) improves over BM25 but does not surpass Run 3 when fusion excludes the dense semantic export from Run 2, which was computed in the notebook but not archived. On unanimity qrels, Run 5 reaches MAP = 0.029, nDCG@1000 = 0.110, and P@10 = 0.078.

4.4. Organizer evaluation

The eRisk 2026 lab overview [1, 2] lists VANGUARD among 12 teams with four submitted runs (45 runs total in the task). Official Task 3 leaderboard metrics were still pending in that overview. Table 2 is our main comparison for now, until final ranks appear in the published proceedings.

5. Discussion and conclusion

Our approach offers a modular hybrid IR grounded in the literature. It combines reproducible sparse and neural baselines with RRF fusion without score calibration across stages. Important limitations remain. The setup is purely zero-shot, exports for the semantic and LLM runs are incomplete, encoding four million sentences is computationally expensive, and ethical constraints limit real-world deployment. If labels become available in future editions, we plan to pursue in-domain fine-tuning, enlarge LLM shortlists, explore ColBERT-style late interaction, and conduct human evaluation of top-ranked sentences.

We presented VANGUARD's zero-shot hybrid retrieval pipeline for the inaugural eRisk Task 3, with four submitted runs and best majority-vote nDCG@1000 of 0.113 on cross-encoder reranking.

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 GPT-4 to perform grammar and spelling checks, paraphrase, and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[3] S. C. Guntuku, J. R. Ramsay, R. M. Merchant, L. H. Ungar, Language of ADHD in adults on social media, Journal of attention disorders 23 (2019) 1475–1485.

[4] L. Chen, J. Jeong, B. Simpkins, E. Ferrara, Exploring the behavior of users with attention-deficit/hyperactivity disorder on Twitter: comparative analysis of tweet content and user interactions, Journal of Medical Internet Research 25 (2023) e43439.

[5] D. Wiechmann, E. Kempa, E. Kerz, Y. Qiao, Transparent but powerful: Explainability, accuracy, and generalizability in ADHD detection from social media data, arXiv preprint arXiv:2411.15586 (2024).

[6] A. Murarka, B. Radhakrishnan, S. Ravichandran, Classification of mental illnesses on social media using RoBERTa, in: Proceedings of the 12th international workshop on health text mining and information analysis, 2021, pp. 59–68.

[7] S. Zanwar, D. Wiechmann, Y. Qiao, E. Kerz, The best of both worlds: combining engineered features with transformers for improved mental health prediction from Reddit posts, in: Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, 2022, pp. 197–202.

[8] Z. Ge, D. Li, Y. Wang, N. Hu, X. Zhu, H. Li, X. Zhang, M. Zhang, S. Qi, Y. Xu, et al., Survey and Experiments on Mental Disorder Detection via Social Media: From Large Language Models and RAG to Agents, arXiv preprint arXiv:2504.02800 (2025).

[9] Y. Zhu, Y. Guo, N. Marchuck, A. Sarker, Y. Wang, Leveraging large language models and traditional machine learning ensembles for ADHD detection from narrative transcripts, arXiv preprint arXiv:2505.21324 (2025).

[10] S. R. Mohammad, F. Kadaei, M. Mohkam, Machine learning approaches to identifying neurodevelopmental disorders using social media data: a systematic review, International Journal of Medical Informatics (2026) 106378.

[11] 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.

[12] D. C. Lohani, V. Chawla, B. Rana, A systematic literature review of machine learning techniques for the detection of attention-deficit/hyperactivity disorder using MRI and/or EEG data, Neuroscience 570 (2025) 110–131.

[13] M. O. Ribas, M. Micai, A. Caruso, F. Fulceri, M. Fazio, M. L. Scattoni, Technologies to support the diagnosis and/or treatment of neurodevelopmental disorders: A systematic review, Neuroscience & Biobehavioral Reviews 145 (2023) 105021.

[14] E. Kerz, S. Zanwar, Y. Qiao, D. Wiechmann, Toward explainable AI (XAI) for mental health detection based on language behavior, Frontiers in psychiatry 14 (2023) 1219479.

[15] 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, pp. 3982–3992.

[16] G. V. Cormack, C. L. 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, 2009, pp. 758–759.

[17] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, arXiv preprint arXiv:2104.08663 (2021).