erisk-cedri at eRisk 2026: MentalBERT with BDI-II Filtering and Empathy Detection for Contextualized Early Depression Detection

Filipa Afonso[1,*], Rui Pedro Lopes[1]

[1]Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract

This paper describes the participation of team erisk-cedri in Task 2 of eRisk 2026, the second edition of the Contextualized Early Detection of Depression task. The task requires sequential binary classification of Reddit users from full discussion threads released round by round. MentalBERT was fine-tuned with Focal Loss on the eRisk 2025 dataset provided by the organizers (909 users, 102 depressed and 807 control), and 264 synthetic users were generated from BDI-II symptom templates to bring the training set to 990 users. At inference time, posts are first scored against 24 BDI-II templates with multilingual-e5-large-instruct, the top 20 most relevant are kept, and a regex-based empathy filter removes posts that express support to others rather than personal distress. Five runs were submitted with decision thresholds from 0.75 to 0.95. Run 0 (threshold 0.75) obtained F1 of 0.75, ERDE50 of 0.07, and a median latency of 1.5 posts with speed 1.00. The system ranked 9th out of 17 teams by F1, but its strongest result was detection latency: it obtained the lowest LatencyTP and the highest speed factor among the 10 teams that processed all 500 users.

Keywords

depression detection, early risk detection, MentalBERT, BDI-II, Focal Loss, empathy filter, eRisk 2026

1. Introduction

The CLEF eRisk lab has been running early risk detection challenges since 2017 [1]. Task 2 of eRisk 2026 [8, 9] is the second edition of the Contextualized Early Detection of Depression task. The main difference from earlier eRisk editions is that systems do not see only the posts written by the target user, but the entire Reddit discussion thread, including posts from all other participants and the reply structure that links them.

Threads are released one at a time. After each thread, the system has to submit a binary decision (depressed or not) and a confidence score before the next thread becomes available. Evaluation combines standard classification metrics (Precision, Recall, F1) with latency-aware measures such as ERDE50 and Flatency [2], which penalize correct decisions that arrive too late.

The proposed system follows the architecture of the eRisk 2025 winner HIT-SCIR [3], which used MentalBERT, BDI-II symptom templates for post filtering, and a memory buffer of the most relevant posts to reach F1 of 0.85. This work adapts that approach with three changes:

1. MentalBERT is fine-tuned with Focal Loss to handle the strong class imbalance in the training data (about 1 to 8 between depressed and control users).
2. A simple template-based data augmentation procedure is used to add synthetic users with subtle depression patterns, instead of LLM-based augmentation.
3. A regex-based empathy filter is added to reduce false positives caused by users who reply to others with supportive language.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the system. Section 4 reports the official results. Section 5 discusses them and Section 6 concludes.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

a61183@alunos.ipb.pt (F. Afonso); rlopes@ipb.pt (R. P. Lopes)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related Work

The eRisk lab has refined its evaluation framework over several editions [1]. ERDE was the original early risk metric, but it has known interpretability issues. Sadeque et al. [2] proposed Flatency, which multiplies the F-score by a speed factor computed from the median latency of true positives. This is the metric used in Task 2 of eRisk 2026.

In the 2025 edition of Task 2, HIT-SCIR [3] obtained the best results with F1 of 0.85 and ERDE50 of 0.03. Their system used three components: contextual data augmentation with Phi-4 to generate synthetic conversation context for training, a BDI-II symptom screening module trained end to end with a Straight-Through Estimator, and a MentalBERT-based detector with a dynamic memory buffer. The second team, ELiRF-UPV (F1 of 0.79), tested long-context models like Longformer and MentalLongformer. The third team, HU [4] (F1 of 0.75), used Llama 3.1 to summarize threads to 200 words and then classified them with BERT-base, combined with a sliding window decision module.

MentalBERT [5] is a BERT model pre-trained on mental health social media data. It has been shown to outperform general BERT in depression-related tasks, and it was used by all top teams in eRisk 2025 Task 2. The Beck Depression Inventory II [6] is a 21-item self-report questionnaire that has been widely used in clinical practice. Its symptoms can be turned into short natural language templates and used as semantic anchors for filtering posts.

3. System Description

3.1. Pipeline Overview

For each user and each new thread released by the server, the system runs the following steps:

1. Extract the posts written by the target user in the thread.
2. Score each post against 24 BDI-II templates using cosine similarity over E5 embeddings, and keep the top 20 with the highest score.
3. Apply the empathy filter and discard posts that match support-language patterns.
4. Concatenate the remaining posts in reverse chronological order (most recent first) and pass them to MentalBERT.
5. Compare the output probability with the threshold of the run and emit a decision and a score.

Once a positive decision is issued for a user in any round, it is kept in all later rounds.

3.2. Training Data

The model was trained on the eRisk 2025 Task 2 dataset, which the organizers released as official training data for the 2026 edition. The dataset contains 909 users (102 depressed, 807 control), with a class imbalance of about 1 to 8. Each user is represented as a sequence of Reddit threads in JSON format, with all posts from all participants and the parent-child reply structure. The dataset was split 80/20 into training (about 726 users) and validation (about 183 users).

3.3. Data Augmentation

The validation analysis of the base model revealed five symptoms with poor performance due to rare-class scarcity in the training data: suicidal thoughts (s9), indecisiveness (s13), punishment feelings (s6), irritability (s11), and self-dislike (s7). To address this imbalance, the training set was augmented with 264 synthetic depressed users targeting these specific symptoms.

The synthetic data was built in two stages. First, an LLM (Claude 3.5 Sonnet) was prompted to generate realistic and varied sentences in the style of Reddit posts for each of the five under-represented symptoms. Second, the generated sentences were consolidated into a static Python list of approximately 750 sentences (around 150 per symptom) through a script (`create_synthetic_data.py`). The script

constructs a DataFrame from this list, where each row contains one sentence and a binary indicator marking the target symptom as positive and all other symptoms (s1–s21) as zero. Each synthetic user is assigned a synthetic identifier (`synth_0`, `synth_1`, ...) and the sentence is duplicated into the `full_text` field. The resulting CSV is concatenated with the real training data prior to fine-tuning.

To improve generalization, the synthetic sentences mix formal register (e.g., "I want to end it all.") with informal Reddit-style register (lowercase text, slang, abbreviations such as "tbh" or "idk"), reflecting the variety of writing styles present in the real corpus. No thread structure was simulated and no synthetic timestamps were assigned: each synthetic user consists only of first-person symptom expressions, without replies from other interlocutors. The synthetic users were mixed exclusively into the training split, leaving the validation split unchanged. The augmented training set therefore contains 990 users (726 original plus 264 synthetic).

3.4. BDI-II Post Filtering

The system uses 24 BDI-II templates: 3 direct depression indicators and 21 indirect symptoms. The direct ones are "I feel depressed", "I am diagnosed with depression", and "I am treating my depression". The indirect ones are short statements such as "I feel sad", "I am discouraged about my future", "I have thoughts of killing myself", "I don't have energy", "I feel worthless", and so on, one per BDI-II item.

The embedding of each post and of each template is computed with multilingual-e5-large-instruct, and cosine similarity is used to score the post against every template. The post score is the maximum similarity across the 24 templates. For each user, the 20 posts with the highest score are kept and passed to MentalBERT.

3.5. Empathy Filter

The validation analysis also revealed two false positives where the misclassified users were not depressed but were replying with supportive language to others who were. Their posts contained phrases like "I'm sorry to hear that", "stay strong", or "sending love". These posts share lexical and semantic content with depression-related language, which makes them hard for the classifier to distinguish.

A simple regex-based filter was added to remove such posts before they reach the classifier. The patterns include:

- Expressions of sorrow for others: "I'm sorry to hear", "I'm sorry for you", "that must be"
- Affirmations of support: "here for you", "sending love", "you're not alone"
- Encouragement: "stay strong", "hang in there"
- Acknowledgement of difficulty: "sounds hard", "sounds rough", "that sucks"
- Physical comfort: "hugs"

Posts that match at least two of these patterns are flagged as empathy posts and excluded from the input to MentalBERT.

3.6. MentalBERT Fine-Tuning

The classifier is mental/mental-bert-base-uncased fine-tuned for binary depression classification. The training configuration is shown in Table 1.

Focal Loss [7] with gamma 2.0 was chosen because it down-weights easy negative examples and forces the model to focus on the harder positive cases. This is useful when the positive class is rare, as is the case here. Class weights were also applied to further balance the loss.

The posts are concatenated in reverse chronological order so that, if the input exceeds 512 tokens and has to be truncated, the most recent content is preserved. The model produces a single probability for the positive class through a linear head with sigmoid activation.

The base model (without augmentation) reached F1 of 0.842 on the validation set (threshold 0.70, Precision 0.941, Recall 0.762, 16 true positives, 1 false positive, 5 false negatives, 161 true negatives).

Table 1
Training hyperparameters for the MentalBERT classifier.

Hyperparameter	Value
Base model	mental/mental-bert-base-uncased
Max sequence length	512 tokens
Top-K posts per user	20
Batch size	16
Learning rate	1e-5
Epochs	10
Early stopping patience	3
Warmup ratio	0.1
Weight decay	0.01
Hidden dropout	0.2
Loss function	Focal Loss (gamma 2.0) with class weights
Train/val split	80/20

Table 2
Decision thresholds used in each run.

Run	Threshold	Strategy
0	0.75	Aggressive
1	0.80	Moderate
2	0.85	Conservative
3	0.90	Strict
4	0.95	Ultra-safe

Table 3
Official decision-based results for erisk-cedri in Task 2 of eRisk 2026.

Run	τ	P	R	F1	ERDE50	LatencyTP	Speed	Flatency
0	0.75	0.73	0.77	**0.75**	**0.07**	**1.5**	**1.00**	**0.75**
1	0.80	0.78	0.51	0.61	0.10	4.5	0.99	0.60
2	0.85	0.85	0.44	0.58	0.11	5.5	0.98	0.57
3	0.90	0.84	0.34	0.48	0.12	9.0	0.97	0.47
4	0.95	1.00	0.11	0.20	0.15	9.0	0.97	0.19

After retraining on the augmented dataset of 990 users, the model reached F1 of 0.8636 on the same validation set.

3.7. Five Submission Runs

The official task allows up to 5 runs. The same model was used in all 5 runs, with only the decision threshold varying, as shown in Table 2. The idea was to cover the precision-recall trade-off and observe where the optimal threshold falls in the test set.

4. Results

4.1. Decision-Based Evaluation

Table 3 shows the official results of the five runs. Run 0 obtained the best F1, ERDE50, and Flatency. The latency of 1.5 posts and the speed of 1.00 mean that half of the correctly detected positive users were detected within the first two posts of their threads.

As the threshold increases, precision rises monotonically from 0.73 to 1.00, but recall drops sharply

Table 4
erisk-cedri Run 0 compared to the other teams with complete coverage (500 user threads), ordered by F1.

Rank	Team	F1	Flatency	LatencyTP	Speed
1	HUGETIME	0.83	0.81	9.0	0.97
2	UNED-GELP	0.82	0.81	4.0	0.99
3	NoirByte	0.82	0.76	20.0	0.93
4	MindAILab	0.82	0.77	16.0	0.94
5	INSA-Lyon	0.80	0.76	13.0	0.95
6	LCDAA	0.78	0.75	10.0	0.96
7	UET-PsyWar	0.78	0.74	15.0	0.95
8	Lotu-ixa	0.76	0.74	6.5	0.98
9	**erisk-cedri**	**0.75**	**0.75**	**1.5**	**1.00**
10	VANGUARD	0.74	0.72	6.5	0.98

Table 5
NDCG@100 for erisk-cedri at different numbers of observed writings.

Run	1 writing	100 writings	250 writings	500 writings
0	0.65	0.81	0.84	**0.84**
1	0.60	0.77	0.77	0.76
2	0.63	0.77	0.76	0.77
3	0.63	0.75	0.73	0.75
4	0.63	0.72	0.71	0.70

from 0.77 to 0.11. The cost of the higher threshold is much larger than the precision gain, and F1 degrades in all later runs.

4.2. Comparison with Other Teams

Table 4 compares Run 0 with the other 9 teams that processed all 500 user threads. The system ranked 9th out of 17 teams when ordered by F1. However, the strongest result of this submission was detection latency: the system obtained the lowest LatencyTP (1.5 posts) and the highest speed factor (1.00) of any complete-coverage submission. The best team by F1 (HUGETIME) reached F1 of 0.83 with a latency of 9 posts, but no team matched this submission's combination of speed=1.00 and competitive F1.

The system is the only one with speed 1.00 in the complete-coverage group. The full 500-user evaluation was completed in 1 hour and 47 minutes, the second fastest after Snugi-AI-v2 (1 hour 26 minutes).

4.3. Ranking-Based Evaluation

Table 5 shows the NDCG@100 values for the five runs at different numbers of observed writings. Run 0 reached NDCG@100 of 0.84 at 500 writings, with P@10 of 0.90 and NDCG@10 of 0.94 at the same point. The values at 1 writing are lower because the system has very little evidence at that point.

5. Discussion

5.1. Low Latency

The most positive aspect of the results is the latency. A LatencyTP of 1.5 posts means that the system detected most positive users very early in their thread sequences. The speed factor of 1.00 is the maximum value in the Flatency framework. In a real monitoring scenario, where the goal is to identify users at risk as soon as possible, this is a useful property. The cost is a moderate Precision of 0.73, which means that about one in four alerts is a false positive.

5.2. The Recall Drop

The five runs show a clear asymmetry between threshold and performance. Going from 0.75 to 0.80 loses 26 points of recall and gains only 5 points of precision. Going to 0.95 reaches perfect precision but only detects 11% of the positive users. This means that most positive users get probability scores between 0.75 and 0.85, not above 0.90. They are borderline cases where the model is not fully confident, probably because their depression is expressed in subtle or indirect ways.

The data augmentation procedure improved performance on the rare-class symptoms targeted at training time, raising validation F1 from 0.842 to 0.8636. However, on the test set the effect did not seem strong enough to lift borderline cases above the 0.80 threshold, suggesting that the gap is also driven by factors beyond rare-symptom coverage. A larger and more diverse synthetic corpus, possibly combined with calibration techniques such as temperature scaling, could help in future work.

5.3. Empathy Filter

An ablation of the empathy filter on the official test set is not available, but on the validation set the filter removed the two false positives that had been identified in the error analysis. The precision of 0.73 on the test set is similar to the validation result, so the filter probably contributed to keeping it at this level. A more systematic evaluation, with and without the filter, is needed to confirm this.

5.4. Gap to the 2025 Winner

The best F1 of 0.75 is 0.08 points below HIT-SCIR's 2025 result (0.85). Three components of their system that were not replicated in this work are likely behind this gap:

1. LLM-based data augmentation with Phi-4 to generate realistic synthetic context, instead of the template-based augmentation used here.
2. A dynamic memory buffer that accumulates the most relevant posts across all rounds, instead of processing each round independently.
3. A voting ensemble of three models trained with 5-fold cross-validation.

Implementing these is a clear direction for the next edition.

6. Conclusion

This paper described the system submitted by team erisk-cedri to Task 2 of eRisk 2026. The system combines MentalBERT fine-tuned with Focal Loss, BDI-II semantic filtering with E5 embeddings, template-based data augmentation, and a regex-based empathy filter. The best run obtained F1 of 0.75, which placed the team 9th out of 17 by F1, while the strongest result was the lowest detection latency (1.5 posts) and the highest speed factor (1.00) among the 10 teams that processed all 500 users.

The main limitation of the system is the recall drop at higher thresholds, which suggests that the model does not produce high-confidence predictions for users with subtle depression. Future work will focus on LLM-based augmentation, cross-round memory mechanisms, and ensemble methods to improve performance on these borderline cases.

Acknowledgments

This work was supported by FCT – Fundação para a Ciência e Tecnologia, I.P. by projects: CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI: 10.54499/UID/PRR/05757/2025); SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: Generate synthetic training data for under-represented symptom classes (as described in Section 3), Drafting content, Paraphrase and reword, and Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5–8, 2016, Proceedings, volume 9822 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 28–39.

[2] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, ACM, 2018, pp. 495–503.

[3] Y. Zi, B. Wang, Y. Zhao, B. Qin, HIT-SCIR@eRisk2025: Exploring the potential of a learnable screening model and risk post buffer-based framework for contextualized early prediction of depression on social media, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9–12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1680–1689.

[4] M. Saad, A. U. Chaudhry, M. Abbas, F. Alvi, A. Samad, Contextualized early detection of depression: Hybrid and time-aware approaches: HU at eRisk Task 2 2025, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9–12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1636–1645.

[5] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, European Language Resources Association, 2022, pp. 7184–7190.

[6] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory-II, Psychological Corporation, San Antonio, TX, 1996.

[7] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, ICCV 2017, IEEE Computer Society, 2017, pp. 2999–3007.

[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, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] 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, Lecture Notes in Computer Science, Springer, 2026.