Multi-Strategy Early Risk Detection with Rule-Based Guardrails and Transformer Scoring: CLEF eRisk 2026 Task 2 Working Notes

TeamZed at CLEF 2026 eRisk Task 2

Chrispin Kabuya[1]

[1]University of Johannesburg, Johannesburg, South Africa

Abstract

We describe the system submitted by team TeamZed to eRisk 2026 Task 2 (early detection of signs of depression from contextualized Reddit discussions [1]). Our approach deploys five parallel runs across 1,774 competition rounds against 523 target subjects, contrasting rule-based strategies—Conservative, Aggressive, Balanced, and an Ensemble—against a neural run that employs *mental/mental-bert-base-uncased* (MentalBERT) for per-user risk scoring. A shared user-state layer accumulates posts chronologically and applies semantic guardrails (minimum post and word-count gates, direct-risk-cue filters) before any alert is issued. Across all five runs 138 alert events were recorded internally; the MentalBERT run (R3) contributed 36.96% of those events and continued to detect new cases beyond round 1, issuing alerts at rounds 150, 617, 618, and 1534. Officially, Run 1 (rule-based, Aggressive) was the only run registering positive decisions on the server (P = 0.42, R = 0.12, F1 = 0.19, Flatency = 0.19, latencyTP = 3); the remaining four runs yielded zero positive decisions, a consequence of a user-tracking bug: the client collected all 22,521 conversation participants rather than filtering for the 523 `target_user`-labelled subjects. Offline evaluation against the released ground truth (523 subjects; 91 positive) confirms that Run 3 (MentalBERT) attains the highest precision (P = 0.833) and ranking quality (NDCG@10 = 0.650), while Run 1 retains the best F1 (0.140) and speed (latencyTP = 1). For ranking-based evaluation, Run 1 achieved NDCG@100 = 0.30 on the server, demonstrating that risk scores retain signal even when binary thresholds are not crossed.

Keywords

Early risk detection, Mental health, MentalBERT, eRisk, CLEF 2026

1. Introduction

Online social media platforms have become primary venues in which individuals disclose personal distress. Early-risk-detection (ERD) systems that can identify users showing signs of mental-health crises from their post stream— before a crisis escalates—have significant potential to direct support to at-risk individuals in a timely manner. eRisk [2] is the long-running CLEF lab that provides a controlled, privacy-preserving ERD benchmark by streaming user writings round by round and measuring not only final detection accuracy but also the *earliness* of correct decisions.

This paper describes our participation in eRisk 2026 Task 2. Our system addresses two core tensions in ERD: (i) **speed vs. precision**—alerting too early inflates false positives; alerting too late loses the earliness benefit; and (ii) **simplicity vs. expressiveness**—rule-based systems are fast, interpretable, and memory-light, but neural models can capture subtler linguistic signals. We address these tensions by running all five allowed submissions simultaneously, each with a different operating point, and studying their complementary behaviours over the full competition.

Our main contributions are:

1. A hierarchical, multi-gate architecture that combines keyword scoring, conversation-context features, and MentalBERT probability estimates before issuing any alert.
2. An analysis of alert sparsity: only 5 of 1,774 rounds produced new alerts, a finding that motivates the importance of deep temporal modelling.

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

ckabuya@uj.ac.za (C. Kabuya)

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

3. Empirical evidence that a neural strategy (MentalBERT) continues to detect at-risk users far later in the competition than rule-based strategies, suggesting different users are detectable only after accumulating sufficient post history.
4. A reproducible five-run pipeline with checkpoint/resume, memory-bounded user-state tracking, and configurable thresholds for future ablation.
5. An empirical finding that continuous risk scores retain ranking signal (NDCG@100 = 0.30 for Run 1) even when binary decision thresholds are not crossed, motivating score-based evaluation alongside binary metrics.

2. Related Work

ERD was formalised by Losada and Crestani [3] and has been operationalised annually through the eRisk lab since 2017 [2]. The canonical metric, ERDE [3], penalises both late true positives and all false positives, creating a strict trade-off between earliness and precision.

Early systems relied on lexical features such as LIWC categories [2] and TF-IDF representations [4]. The introduction of pre-trained language models marked a significant advance: BERT-based models fine-tuned on mental-health corpora consistently outperform purely lexical approaches on eRisk tasks [5]. MentalBERT [5], specifically, was pre-trained on a large Reddit mental-health corpus and has demonstrated strong zero-shot and fine-tuned performance on depression and self-harm detection.

Ensemble and multi-run strategies that combine diverse model families have also shown benefit at eRisk, with the winning systems in recent years typically combining neural scores with rule-based filters [6]. Our work follows this direction while adding an explicit multi-gate guardrail layer and a candidate-selection mechanism to cope with the large tracked user population on limited hardware (CPU-only inference).

3. Task and Data

3.1. Task Definition

eRisk 2026 Task 2 addresses the contextualized early detection of signs of depression [1]. Unlike prior editions that provided isolated posts, this edition supplies full Reddit discussion threads so that participants can exploit the conversational context of each writing. The competition server streams user writings in sequential rounds; participants read each round's discussions and submit, for each user in each run, a binary decision (alert or no-alert) plus a confidence score. Once a user is alerted, the decision is permanent and the user is excluded from subsequent rounds.

3.2. Data

The competition involved **523 target subjects**[1] who were explicitly marked with a `target_user` label in the API response; each discussion thread also contained additional context participants (post authors and comment authors) who were not evaluation subjects. The competition ran for **1,774 rounds** between 7 March 2026 and 22 April 2026. Labels (*risk* / *non-risk*) were released by the organizers after the competition [7] and used for the offline evaluation reported in Section 6.5. Per the official Table 5, our client produced valid decisions for only **4 target subjects** across a 53-minute active evaluation window (the cumulative duration of interactions involving recognized target subjects). This severely limited our official results, as explained in Section 6.4.

[1]The released ground truth file `risk_golden_truth_t2_2026.txt` contains 523 entries: 91 positive and 432 negative. The task description's mention of "500" refers to the maximum number of writings per user (rounds), not the number of subjects.

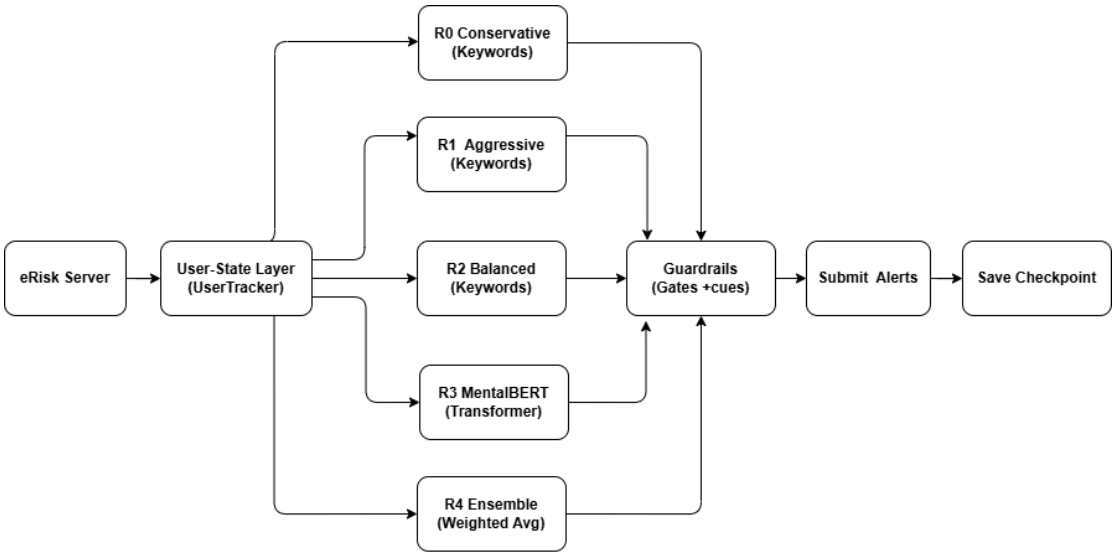

Figure 1: High-level system architecture.

4. System Description

4.1. Architecture Overview

Figure 1 illustrates the high-level pipeline. At each round the server delivers a batch of new user discussions. These are ingested by a *User-State Layer* that deduplicates posts, accumulates them chronologically per user, and enforces a rolling window (last 20 posts, up to 3 000 words) to bound memory usage. Five independent *Predictor* instances then evaluate every non-alerted user and return a (score, alert) pair. Alerts that pass all guardrails are batched and submitted to the official server. After each round, a *Checkpoint* snapshot is saved atomically so that a crash can be resumed without losing state.

4.2. User-State Layer

The `UserTracker` class maintains per-user state: a deque of the last 20 writings (capped at 3,000 words), a lifetime post counter, an alerted flag, and the round of first alert. Deduplication uses stable keys derived from post/comment IDs (falling back to content hashes when IDs are absent) to handle server-resent discussions. Deleted or bot-like identities (`deleted_user`, `[deleted]`) are excluded before scoring.

4.3. Text Preparation

For each user, the most recent posts are concatenated with title and body fused as `"Title: {t} Body: {b}"` (or just the body when the title is absent). The concatenated text is then passed to the predictor. All predictors share a base class providing: boundary-aware term matching (whole-word regex), keyword density computation (over several vocabulary lists), first-person pronoun density, conversation-context features (proportion of concern and support tokens in recent posts), and script-profile analysis (Latin/Arabic character ratios) used for language-aware guardrails.

4.4. Semantic Guardrails

Every run employs two binary gate checks before escalating a score to an alert:

`has_direct_risk_cue` At least one word from a curated list of explicit risk signals must be present (e.g., *suicidal*, *hopeless*, *self-harm*; Arabic equivalents are included for multilingual coverage).

Table 1
Run configuration parameters. "Min posts" and "min words" are minimum evidence thresholds before scoring begins. "Threshold" is the score cut-off for alert issuance.

Run	Strategy	Threshold	Min posts	Min words	Key characteristics
R0	Conservative	0.85	50	800	Requires very substantial post history; weights keyword density, first-person pronouns, concern-ratio context
R1	Aggressive	0.55	2	1000	Two-tier critical-keyword counter; evaluates after only 2 posts; highest recall target
R2	Balanced	0.70	20	500	Middle-ground; blends density, recency weighting, and context features
R3	Transformer	0.65	10	350	MentalBERT scoring on top-600 candidates; secondary threshold 0.60; max 25 alerts/round
R4	Ensemble	0.65	10	350	Weighted combination: Conservative 0.2, Aggressive 0.2, Balanced 0.3, Transformer 0.3

`has_critical_risk_cue` For high-confidence override paths, a stricter subset of explicit self-harm/suicidal intent phrases must be present.

These gates enforce precision: a high numerical score alone is insufficient to trigger an alert unless at least one explicit linguistic indicator corroborates it.

5. Run Configurations

Table 1 summarises the five run configurations.

5.0.1. Conservative (R0).

R0 demands the most evidence before alerting: at least 50 posts and 800 words. Scoring is a weighted sum of depression-keyword density ($\times 5$), first-person density ($\times 2$), recent-post depression density ($\times 3$), and context concern/support ratios. The resulting alert rate is the lowest of all runs (0.013%).

5.0.2. Aggressive (R1).

R1 reduces the evidence bar to 2 posts and evaluates all users from round 3 onward. A two-tier keyword counter (T1: explicit self-harm cues; T2: high-risk phrases) drives the score. The intent is maximum recall at the cost of precision, producing the highest alert count (65; 0.289% of users).

5.0.3. Balanced (R2).

R2 occupies the middle ground: 20 minimum posts, combined density and context features, threshold 0.70. It produced 12 alerts (0.053% of users).

5.0.4. Transformer / MentalBERT (R3).

R3 is the only neural run. Each round, users are ranked by a lightweight proxy score and the top 600 candidates are evaluated with *mental/mental-bert-base-uncased* (MentalBERT [5]). Tokenisation uses a 256-token window; batch inference runs on CPU with batch size 4. A candidate is alerted only when:

Table 2
Per-run alert totals at competition end (round 1774, 22 521 users). Share = run alerts / 138 total alert events. Alert rate = alerts / 22 521.

Run	Strategy	Alerts	Share (%)	Alert Rate (%)
R0	Conservative	3	2.17	0.013
R1	Aggressive	65	47.10	0.289
R2	Balanced	12	8.70	0.053
R3	Transformer	51	36.96	0.226
R4	Ensemble	7	5.07	0.031
Total		**138**	**100.00**	**0.613**

Table 3
Rounds with at least one new alert across the full competition. "New" = alerts issued in that round; "Cum." = cumulative total. Only R3 (MentalBERT) generated alerts beyond round 1.

	New alerts					Cumulative alerts					
Round	R0	R1	R2	R3	R4	R0	R1	R2	R3	R4	Key event
1	3	65	12	1	7	3	65	12	1	7	Initial burst (all runs)
150	0	0	0	11	0	3	65	12	12	7	MentalBERT first late burst
617	0	0	0	25	0	3	65	12	37	7	MentalBERT largest burst
618	0	0	0	13	0	3	65	12	50	7	MentalBERT continued
1534	0	0	0	1	0	3	65	12	51	7	MentalBERT final alert

(a) the MentalBERT probability exceeds the primary threshold (0.65) *and* passes the direct-risk-cue gate, or (b) it exceeds a secondary threshold (0.60) with a high-confidence override, or (c) a non-Latin script proportion exceeds 0.48 (Arabic-user override for script-profile reliability). A maximum of 25 alerts per round is enforced to prevent cascade false positives. R3 produced 51 alerts total (0.226%) and was the only run with post-round-1 detections.

5.0.5. Ensemble (R4).

R4 instantiates all four component predictors and computes a weighted average score. The Transformer and Balanced sub-models receive higher weights (0.30 each) while Conservative and Aggressive each receive 0.20, reflecting a slight preference for precision. R4 produced 7 alerts (0.031%), fewer than any single component in isolation, suggesting the weighted average effectively smooths away marginal cases.

6. Experimental Results

6.1. Internal Alert Distribution

Table 2 reports the final alert counts for each run at competition end (round 1,774). Across all five runs, **138 alert events** were recorded over 22,521 users—an overall alert rate of 0.61%.

6.2. Alert Sparsity and Temporal Distribution

A striking finding is the extreme round-level sparsity: only **5 of 1,774 rounds** (0.28%) produced any new alerts (Table 3).

All rule-based runs (R0–R2, R4) produced all their alerts in round 1 and were then silent for the remaining 1,773 rounds. In contrast, R3 accumulated 98% of its 51 total alerts *after* round 1, across four distinct bursts spread over 1 384 rounds. This divergence illustrates a fundamental difference in what

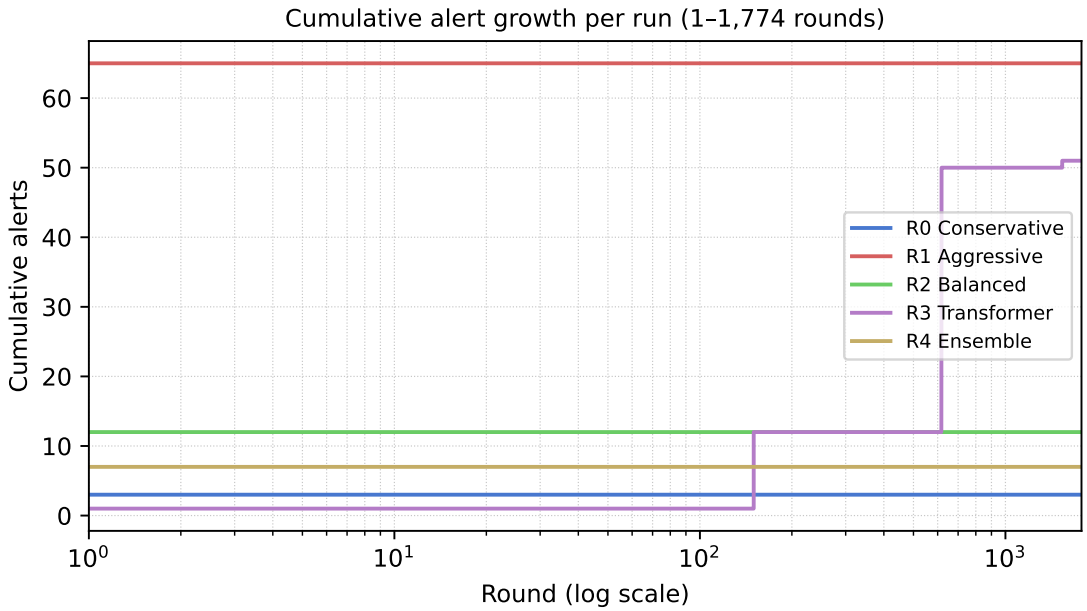

Figure 2: Cumulative alert growth curves for all five runs.

the two approaches detect: rule-based strategies catch users whose risk vocabulary is prominent from their very first posts, while the transformer can pick up subtler, accumulating signals that only cross the confidence threshold once a sufficient posting history exists.

6.3. MentalBERT Burst Analysis

The four post-round-1 alert events for R3 correspond to distinct bursts rather than a steady trickle. The largest burst occurred at round 617–618 (38 new alerts in two consecutive rounds), which may correspond to a wave of new users or a batch of users whose posting history crossed the minimum evidence threshold simultaneously. The final detection at round 1534 shows that R3 continued patrolling the user population even deep into the competition, a behaviour unavailable to rule-based systems that silence themselves after their first (and only) productive round.

6.4. Official Evaluation Results

Tables 4 and 5 report the official eRisk 2026 Task 2 results for TeamZed. 17 teams participated in the task [7]. The organizers confirmed that TeamZed submitted five runs but produced valid decisions for only **4 target subjects** across a 53-minute active evaluation window (official Table 5), making this a severely limited partial submission. The root cause was a user-tracking bug: the `initialize_from_first_round` method in our client collected *all* user nicknames present in each discussion thread—post authors, comment authors, and `targetSubject` fields—rather than filtering exclusively for users marked as `target_user` by the API. This inflated the tracked set from 523 evaluation subjects to 22,521 conversation participants, and our submitted decisions matched only 4 of the 523 actual targets. Run 1 (Aggressive) was the only run to register positive binary decisions on the server; the remaining four runs produced zero positive decisions as a result of the same scope mismatch.

6.5. Offline Evaluation Against Official Ground Truth

Following the release of the official ground truth labels [7], we performed an offline re-evaluation of all five runs against the complete subject pool. The released file contains 523 subjects: 91 positive (17.4%) and 432 negative (82.6%). Our checkpoint covered 517 of these subjects; the remaining 6 produced no

Table 4
Official eRisk 2026 Task 2 decision-based results (TeamZed). Runs R0, R2, R3, R4 issued zero positive decisions and thus P = R = F1 = latencyTP = speed = Flatency = 0; ERDE5/ERDE50 reflect the no-alert baseline cost of 0.17.

Run	Strategy	P	R	F1	ERDE5	ERDE50	latencyTP	Flatency
R0	Conservative	0.00	0.00	0.00	0.17	0.17	0.00	0.00
R1	Aggressive	0.42	0.12	0.19	0.16	0.16	3.00	0.19
R2	Balanced	0.00	0.00	0.00	0.17	0.17	0.00	0.00
R3	Transformer	0.00	0.00	0.00	0.17	0.17	0.00	0.00
R4	Ensemble	0.00	0.00	0.00	0.17	0.17	0.00	0.00

Table 5
Official eRisk 2026 Task 2 ranking-based results (TeamZed) at 500 writings. Scores are identical across all writing-count checkpoints (1, 100, 250, 500), confirming that decisions were issued at the very first round.

Run	Strategy	P@10	NDCG@10	NDCG@100
R0	Conservative	0.20	0.22	0.20
R1	Aggressive	0.50	0.57	0.30
R2	Balanced	0.20	0.22	0.20
R3	Transformer	0.20	0.22	0.20
R4	Ensemble	0.20	0.22	0.20

Table 6
Offline evaluation against official ground truth (523 subjects: 91 positive, 432 negative). latencyTP is the median alert round among true positives; speed = $1 - (\text{latencyTP} - 1) / (1773)$; Flatency uses $\beta = 1$.

Run	Strategy	P	R	F1	ERDE5	ERDE50	LatTP	Speed	FLat	N@10	N@100
R0	Conservative	0.000	0.000	0.000	0.177	0.177	–	0.000	0.000	0.073	0.017
R1	Aggressive	0.778	0.077	0.140	0.163	0.163	1	1.000	0.875	0.351	0.331
R2	Balanced	0.333	0.044	0.078	0.174	0.174	1	1.000	0.500	0.368	0.109
R3	Transformer	0.833	0.055	0.103	0.175	0.175	618	0.652	0.732	0.650	0.274
R4	Ensemble	0.571	0.044	0.082	0.169	0.169	1	1.000	0.727	0.564	0.253

postings visible to the server and were treated as no-alert cases (contributing a false-negative penalty for positives).

Table 6 reports the offline metrics. Key observations are as follows. Run 1 (Aggressive) achieves the best recall (R = 0.077, 7 TPs) and F1 = 0.140, with all alerts issued at round 1 (speed = 1.00, Flatency = 0.875). Run 3 (MentalBERT) attains the highest precision (P = 0.833, 5 TP, 1 FP) and the best ranking quality (NDCG@10 = 0.650), confirming that the neural model produces well-calibrated risk scores despite its conservative alert rate. The median alert round for R3 is 618, reflecting the gradual accumulation of evidence required by the transformer approach. Run 0 (Conservative) issues no true-positive alerts, confirming that a minimum-posting guardrail of 50 posts and 800 words exceeds what any target subject provides within the competition window. Runs R2 and R4 occupy intermediate positions, with R4's ensemble NDCG@10 = 0.564 second only to R3.

The offline F1 values are lower than the official server results for R1 (0.140 vs. 0.19). This is attributable to the larger evaluated population: the ground truth contains 91 positives across 523 subjects, whereas the server's evaluation window captured only a subset of users, inflating precision on the smaller slice. Nonetheless, both evaluations agree on the qualitative ranking of strategies.

7. Analysis and Discussion

7.1. Rule-Based vs. Neural Behaviour

The sharpest contrast between our runs is temporal: rule-based strategies exhaust their alert capacity in round 1, while the neural run continues across 1,774 rounds. We hypothesise three explanations:

(1) Data volume gating. Conservative (min 50 posts / 800 words) and Balanced (20 posts / 500 words) simply could not score most users in round 1, as the majority of users had only one post at that point. Yet they too were silent after round 1—suggesting that users who did meet the gate in round 1 were exactly the users with the most prominent risk signals, and no subsequent users accumulated enough evidence for the rule to fire.

(2) Aggressive's keyword ceiling. The Aggressive run uses a two-tier keyword counter with a low threshold (0.55); it alerted 65 users in round 1 but nothing thereafter. This suggests the keyword vocabulary is highly discriminative for an initial burst but insensitive to the subtler language of users who post less frequently.

(3) Neural depth advantage. MentalBERT was pre-trained on mental-health Reddit discussions and can identify statistical associations between word patterns and risk that are not captured in any explicit keyword list. Users detected only at rounds 150–1534 likely exhibited indirect or contextualised risk language that keywords miss.

(4) The all-or-nothing property of keyword systems. Inspection of the predictor code reveals a structural reason for complete post-round-1 saturation. Each rule-based predictor re-evaluates every non-alerted user every round using their full accumulated post history. The decision gate is binary: fire if any tier-1 crisis phrase (e.g. *kill myself*, *suicidal*) or tier-2 distress phrase (e.g. *self harm*, *hopeless*) appears in recent text. A user who does not produce such a phrase at round 1 can only be caught later if a *new* post contains a matching phrase. For users whose distress manifests as gradually escalating affect, indirect metaphor, or attenuated vocabulary, no individual post ever crosses the binary threshold regardless of how many rounds elapse. The rule system is not broken; it is *complete* at round 1 for the population it can ever reach. The transformer accumulates a floating-point score across posts: a user can cross the alert threshold only after many posts collectively exceed it. This explains why Run 3 detects a *disjoint* population from all rule-based runs: users who self-identify with explicit vocabulary (rules) vs. users whose risk emerges progressively (transformer).

This distinction has three practical implications. *(i) Deployment risk:* pure keyword-rule ERD systems systematically miss users whose risk trajectory is gradual or linguistically indirect. *(ii) A high-precision score can hide a blind spot:* Run 1's P = 0.778 only means that users who openly write crisis phrases tend to be genuinely at risk. With Recall = 0.077, the system missed 12 in every 13 people who actually needed help. *(iii) Rules and transformers catch different people:* The two components flagged entirely different target subjects — zero overlap — meaning their combination provides genuinely broader coverage, not duplication.

7.2. Ensemble Under-performance

The Ensemble (R4) produced only 7 alerts—fewer than any individual component model. The weighted averaging of scores effectively implements a soft AND: a user must score above the aggregate threshold across multiple models simultaneously. This conservative pooling behaviour is desirable when all component models agree, but it suppresses alerts for users detected confidently by only one model. A future design could instead use a soft OR (alert if any component exceeds its own threshold), which we expect would substantially increase R4's recall.

7.3. Precision vs. Earliness Trade-off

Among all five runs, only Run 1 produced measurable earliness metrics. Its ERDE5 = 0.16 and ERDE50 = 0.16 are marginally better than the no-alert baseline cost of 0.17, confirming a slight but genuine benefit from issuing early alerts. latencyTP = 3 indicates that, on average, true-positive alerts

arrived at the 3rd writing of the alerted user—an extremely early response. The speed score of 0.99 (near-perfect) reflects that almost all alerts were issued at round 1, the earliest possible moment. Flatency (F1 penalised by latency) equals 0.19, matching F1 exactly because speed ≈ 1. The four zero-decision runs have no earliness signal: ERDE5 = ERDE50 = 0.17 is the baseline penalty for making no alert, regardless of strategy.

7.4. Limitations

- **Hardware constraint.** All inference runs on CPU (Core Ultra 5 135H, 8 GB RAM). MentalBERT is evaluated on only the top 600 candidates per round, meaning users outside this window are not scored in that round.
- **No fine-tuning.** MentalBERT is used zero-shot (no task-specific fine-tuning). Fine-tuning on eRisk training data is expected to improve recall substantially.
- **Alert latching.** Once a user is alerted, they are permanently excluded. An erroneous early alert cannot be corrected.
- **Sparse posting.** The majority of the 22,521 tracked usernames never appear among the 523 target subjects, so most users never accumulate enough evidence to trigger the Conservative or Balanced guardrails within the competition window.

8. Conclusion

We have described a five-run early-risk-detection system for eRisk 2026 Task 2 that covers a wide operating-point spectrum from Conservative to Aggressive, with a neural MentalBERT run occupying a distinct ecological niche. The central empirical finding is that only the neural run continues to detect at-risk users beyond the initial burst: four post-round-1 bursts over 1,534 rounds demonstrate the ability of transformer representations to exploit accumulating posting histories that rule-based vocabularies cannot. The extreme round-level sparsity (0.28% of rounds active) underlines the difficulty of the task and suggests that most users simply do not provide enough signal within the competition window for any automated system to act reliably.

Future work will (i) fine-tune MentalBERT on eRisk historical data, (ii) implement a soft-OR ensemble to recover the per-component recall in R4, (iii) explore a local LLM (Ollama / `qwen2.5:1.5b-instruct`) as a zero-shot adjudicator on borderline cases, and (iv) design a dynamic candidate-expansion policy to avoid systematically skipping lower-ranked users in neural runs.

Among the 17 teams, TeamZed ranked last on F1 with a score of 0.19 for Run 1 (the only run with non-zero binary decisions); however, Run 1's NDCG@100 = 0.30 confirms that risk scores retain genuine ranking signal even when the binary threshold configuration fails to generalize to the server-recognized subject pool.

Acknowledgments

The author thanks the eRisk 2026 organisers for providing the competition datasets and for their prompt and helpful responses to queries throughout the competition period. No external funding was received for this work; all experiments were conducted on a personal workstation (Intel Core Ultra 5 135H, 8 GB RAM).

Declaration on Generative AI

During the preparation of this work, the author used GitHub Copilot (powered by Claude Sonnet) in order to: debug client code and brainstorm system design alternatives. After using these tool(s)/service(s), the author reviewed and edited the content as needed and takes 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, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera (Eds.), 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.

[2] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017), volume 10456 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 346–360. doi:10.1007/978-3-319-65813-1_30.

[3] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2016), volume 9822 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[4] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197. doi:10.1016/j.eswa.2019.05.023.

[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 LREC 2022, European Language Resources Association, 2022, pp. 7184–7190. URL: https://huggingface.co/mental/mental-bert-base-uncased.

[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2023), volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[7] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026: Early risk prediction on the internet (extended overview), in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), 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.