Ensemble of Zero-Shot Classifiers with Lexical Boosting for Early Symptom Detection in Therapeutic Conversations

Jose Alberto Pacheco Serna[*,†], Jairo Enrique Serrano Castañeda[†], Edwin Puertas[†] and Juan Carlos Martínez-Santos[†]

Universidad Tecnologica de Bolivar, Cartagena, Colombia

Abstract

Early identification of mental health risks from natural dialogue is a critical challenge in computational psychiatry. This paper describes the system architecture developed for the MentalRiskES task at IberLEF 2026, focusing on the early detection of mental health symptoms in therapeutic conversations using an item-level multiclass classification strategy. Operating under a strict zero-shot constraint where no labeled training data are available, our approach leverages a heterogeneous ensemble of pretrained Transformer models including mDeBERTa-v3, XLM-RoBERTa, and a native Spanish BERT variant. The system integrates context-aware sequential history windows, a multi-template hypothesis projection layer, and a rule-based lexical boosting mechanism targeting specific clinical indicator tokens. To accommodate varied evaluation trade-offs, three execution profiles (Reactive, Hybrid, and Historical) are evaluated using flexible calibrated threshold mapping functions across the GAD-7, PHQ-9, and CompACT-10 psychometric questionnaires. Our experimental results show that the Reactive profile optimized anxiety tracking with a GAD-7 MAE of 0.800, while the Historical profile stabilized long-term predictions for depression, yielding our best PHQ-9 MAE of 1.005. Furthermore, computational efficiency and carbon footprints are audited using local emissions tracking tools.

Keywords

MentalRiskES, IberLEF 2026, Zero-Shot Classification, Ensemble Methods, Lexical Boosting, Computational Psychiatry

1. Introduction

Automated mental health assessment via computational linguistics bridges computer science and clinical psychology. NLP architectures have evolved rapidly from basic keyword matching to deep learning frameworks capable of capturing subtle semantic cues in human speech. The MentalRiskES shared task at IberLEF 2026 addresses this challenge, focusing on early symptom identification within multi-turn therapeutic dialogues [1, 2].

Unlike standard text classification benchmarks that assign static, document-level labels to isolated essays or posts, this task requires sequential evaluation. Systems must process patient-therapist interactions turn by turn across consecutive conversational rounds. After each patient response, the framework maps unstructured natural language to structured, multi-class psychometric outputs that mirror standardized clinical diagnostic instruments.

A key constraint of the MentalRiskES task is the total lack of domain-specific labeled training data. This rules out standard supervised fine-tuning, requiring robust zero-shot, weakly supervised, or knowledge-guided translation strategies instead. Additionally, because emotional states and psychiatric symptoms evolve over time, the system must balance immediate linguistic cues from the current turn against the cumulative context of preceding exchanges.

IberLEF 2026, September 2026, León, Spain

[*]Corresponding author.

[†]These authors contributed equally.

sernaj@utb.edu.co (J. A. P. Serna); jserrano@utb.edu.co (J. E. S. Castañeda); epuertas@utb.edu.co (E. Puertas); jcmartinez@utb.edu.co (J. C. Martínez-Santos)

0009-0006-3903-2028 (J. A. P. Serna)

To address these challenges, we introduce a modular zero-shot ensemble framework optimized for sequential conversational inference. The system relies on three main components:

- A heterogeneous ensemble of multilingual and Spanish-native Natural Language Inference (NLI) models that project raw utterances into formal psychometric dimensions.
- A dual-stream history aggregation layer that separates immediate conversational signals from a dynamic, localized historical baseline window.
- A deterministic, rule-based lexical boosting layer designed to adjust raw probability distributions when high-acuity psychiatric indicator tokens are observed.

We evaluate the system performance across three distinct runtime configurations (`reactive`, `hybrid_es`, and `historical`) to analyze the trade-offs between sensitivity and precision under different decision thresholds. Finally, we audit the entire pipeline for computational and environmental efficiency using carbon tracking tools to assess its viability for real-time deployment.

2. Task Description and Psychometric Instruments

The MentalRiskES challenge frames symptom tracking as an item-level multiclass classification task over sequential interactions. For each active user session, the system processes text streams iteratively and returns ordinal scores matching three clinical questionnaires:

- **GAD-7 (Generalized Anxiety Disorder-7):** This screening tool consists of 7 items designed to gauge the severity of generalized anxiety symptoms. Each item maps to a 4-point Likert scale ranging from 0 ("Not at all") to 3 ("Nearly every day").
- **PHQ-9 (Patient Health Questionnaire-9):** A 9-item instrument based on the DSM-IV diagnostic criteria for depressive disorders. Items are mapped to an ordinal 4-point Likert scale (0 to 3), covering areas such as anhedonia, sleep disturbances, fatigue, and suicidal ideation.
- **CompACT-10 (Comprehensive Assessment of Acceptance and Commitment Therapy processes):** A 10-item questionnaire that tracks psychological flexibility, mindfulness, and behavioral avoidance. This instrument uses a 7-point scale (0 to 6), requiring finer classification boundaries.

The evaluation framework operates as a streaming pipeline. The evaluation server provides a JSON object containing the user's raw text and the current round indicator. The system must immediately return predictions across all 26 dimensions (7 for GAD-7, 9 for PHQ-9, and 10 for CompACT-10) before the server transmits the next conversational turn.

3. System Architecture

The architecture is structured as a deterministic pipeline that transforms raw, unstructured text into calibrated psychometric indices. The workflow consists of four stages: contextual text formatting, multi-model zero-shot NLI projection, cross-stream fusion with trend and lexical adjustments, and threshold-based quantization.

3.1. Contextual State Preservation and Data Ingestion

To handle multi-turn dialogues without performance degradation or excessive context lengths, the system maintains a session-specific text buffer. For any incoming patient input U_t at round t within session S, the utterance is appended to a history array:

$$\text{History}(S) = [U_1, U_2, \ldots, U_t] \tag{1}$$

During evaluation, we formulate two distinct text streams to guide the inference engines:

1. **Current Context Stream (T_{current}):** Captures the immediate state from the latest turn. It truncates the raw string U_t to the first 2000 characters to prevent overly long inputs:

$$T_{\text{current}} = \text{Truncate}(U_t, 2000) \tag{2}$$

2. **Historical Context Stream (T_{history}):** To avoid context drift from long histories, we implement a sliding window that concatenates the four preceding utterances, excluding the current turn, up to 2000 characters:

$$T_{\text{history}} = \text{Truncate}\left(\bigcup_{i=\max(1,t-4)}^{t-1} U_i, 2000 \right) \tag{3}$$

3.2. Heterogeneous Zero-Shot NLI Ensemble

The core engine reframes item classification as a Natural Language Inference (NLI) task. Given a text stream T and a psychometric label $l \in L$, the models estimate the probability that T entails a hypothesis $H(l)$ generated by embedding l into a template τ.

To reduce template bias, we evaluate predictions across multiple prompt formulations. Depending on the runtime profile, these templates cover three distinct perspectives:

- **Clinical:** Focuses on observation markers, e.g., *"Durante las últimas semanas, la persona presenta {l}."*
- **Neutral:** Employs standard framing, e.g., *"La persona muestra {l}."*
- **Strict:** Asserts absolute clinical presence, e.g., *"Existe presencia clara de {l}."*

The ensemble relies on three transformer backbones pretrained on diverse multilingual and localized NLI datasets:

- **Model A (XLM-RoBERTa-Large-XNLI):** A cross-lingual model optimized via multi-task objectives (`joeddav/xlm-roberta-large-xnli`), providing strong generalization across diverse syntactic layouts.
- **Model B (mDeBERTa-v3-base-mnli-xnli):** An architecture integrating disentangled attention mechanisms (`MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`), well-suited for zero-shot boundary resolution.
- **Model C (BERT-Base-Spanish-WWM-XNLI):** A native Spanish model trained with whole-word masking (`Recognai/bert-base-spanish-wwm-cased-xnli`) to better capture regional syntax and idiomatic expressions.

For a given stream T, the ensemble average probability score for a targeted item label l under a set of models $\mathcal{M}$ and template set $\mathcal{T}$ is computed as:

$$P(T, l) = \frac{1}{|\mathcal{M}| \cdot |\mathcal{T}|} \sum_{m \in \mathcal{M}} \sum_{\tau \in \mathcal{T}} \text{Classifier}_m(T, \tau(l)) \tag{4}$$

This operation is executed independently for both streams, yielding $p_{\text{curr}} = P(T_{\text{current}}, l)$ and $p_{\text{hist}} = P(T_{\text{history}}, l)$.

3.3. Lexical Boosting and Trend Modulation

To ensure that explicit, high-acuity clinical markers are not obscured by the soft-max attention mechanisms of the transformer backbones, a deterministic lexical boosting module runs in parallel. We define a targeted set of diagnostic keyword roots M:

$$M = \{\text{suicid}, \text{muerte}, \text{ansiedad}, \text{panico}, \text{depres}, \text{culpa}, \text{inutil}, \text{insomnio}, \text{agotad}, \text{sin salida}\}$$

The system counts unique keyword matches in the lowercase text. A linear boosting coefficient β is computed and bounded by a ceiling to prevent score explosion:

$$\beta = \min\left(0.08, \sum_{m \in M} \mathbb{I}(m \in T_{\text{current}}) \times 0.01\right) \tag{5}$$

The constant increment of 0.01 per keyword and the global saturation cap of 0.08 were heuristically determined during preliminary alpha testing to ensure that explicit critical crisis disclosures inject a significant proportional shift (up to 8% of the probability mass) without completely overriding the semantic context captured by the NLI models.

Additionally, the system monitors emotional acceleration over time. If the current probability p_{curr} exceeds the historical baseline p_{hist} by a margin ($\Delta > 0.12$), a trend adjustment $\Delta_{\text{trend}} = 0.03$ is applied. This threshold gap ($\Delta > 0.12$) was selected to filter out minor conversational fluctuations, isolating true emotional escalations. The final score p_{final} is a weighted combination of both streams:

$$p_{\text{final}} = \min\left(1.0, \max\left(0.0, w_{\text{curr}} \cdot p_{\text{curr}} + w_{\text{hist}} \cdot p_{\text{hist}} + \Delta_{\text{trend}} + \beta\right)\right) \tag{6}$$

where $\Delta_{\text{trend}} = 0.03$ if $p_{\text{curr}} > p_{\text{hist}} + 0.12$, else $\Delta_{\text{trend}} = 0.0$.

3.4. Threshold Calibration and Quantization

The final step maps the continuous score $p_{\text{final}} \in [0, 1]$ into the discrete integers required by each questionnaire. For questionnaires with a maximum score of 3 (GAD-7 and PHQ-9), a 3-element vector $\Theta_3 = [\theta_1, \theta_2, \theta_3]$ is used. For the 7-point CompACT-10, a 6-element vector $\Theta_6 = [\theta_1, ..., \theta_6]$ is applied. The discrete score S is computed as:

$$S = \max\{i \in \{0, \ldots, K\} \mid p_{\text{final}} \geq \theta_i\} \tag{7}$$

with $\theta_0 = 0.0$, and where K denotes the maximum possible scale score.

Given the zero-shot nature of the task and the total absence of a development set or labeled training metrics, these threshold boundaries could not be tuned using standard validation algorithms. Instead, we performed an analytical calibration by mapping the expected probability density function of an NLI model under cross-entropy constraints into uniform and slightly skewed partitions of the $[0, 1]$ interval. The *Sensitive*, *Balanced*, and *Conservative* strategies represent safe margins modeled on the theoretical distribution of Likert intensities, serving as an operational proxy for empirical tuning data.

4. Experimental Configurations and Run Profiles

To evaluate the trade-offs between immediate response sensitivity and long-term tracking, we define three execution profiles. Each profile uses distinct model combinations, weight distributions, and quantization thresholds, as summarized in Table 1.

Table 1
Architectural Specification of System Run Profiles

Profile Name	Active Model Ensemble	w_{current}	w_{history}	Threshold Strategy
`reactive`	XLM-RoBERTa, mDeBERTa	0.80	0.20	Sensitive
`hybrid_es`	XLM-RoBERTa, Spanish-BERT	0.70	0.30	Balanced
`historical`	XLM-RoBERTa, mDeBERTa	0.55	0.45	Conservative

4.1. Run Profile Taxonomy

1. **Reactive Profile (`reactive`):** Prioritizes rapid responses to sudden emotional shifts. It places higher weight on the immediate turn ($w_{\text{curr}} = 0.80$, $w_{\text{hist}} = 0.20$) using an ensemble of XLM-RoBERTa and mDeBERTa with a permissive threshold configuration ($\Theta_3 = [0.18, 0.38, 0.62]$).
2. **Hybrid Spanish Profile (`hybrid_es`):** Incorporated to leverage native Spanish language patterns, pairing the Spanish BERT model with XLM-RoBERTa. It balances the stream weights ($w_{\text{curr}} = 0.70$, $w_{\text{hist}} = 0.30$) and uses intermediate quantization steps ($\Theta_3 = [0.24, 0.47, 0.70]$).
3. **Historical Profile (`historical`):** Focuses on long-term emotional trends while minimizing transient noise. It increases the influence of past context ($w_{\text{hist}} = 0.45$) and uses a conservative quantization layout ($\Theta_3 = [0.30, 0.56, 0.79]$), requiring sustained high probabilities to trigger higher severity scores.

4.2. Computational Infrastructure and Sustainability Audit

Following sustainable AI practices, all experiments were conducted locally to accurately track hardware energy consumption. The execution environment consists of a desktop equipped with an AMD Ryzen 5 9600X processor, 32GB of RAM, and an NVIDIA GeForce RTX 5070 Ti GPU with 16GB of RAM.

The evaluation scripts, built using PyTorch and Hugging Face, run alongside a tracking daemon managed by the `CodeCarbon` package. During execution, the tracker records real-time power draw from the GPU core, CPU sockets, and RAM, logging the execution duration and estimated carbon emissions. This experimental setup allows us to evaluate the accuracy of our zero-shot architectures alongside their ecological impact.

5. Results and Discussion

The official evaluation metrics for the MentalRiskES task 1 provide an objective baseline to analyze how different runtime configurations affect item-level symptom tracking. Table 2 summarizes the Mean Absolute Error (MAE) scores achieved by the three configurations across the GAD-7, PHQ-9, and CompACT-10 scales, alongside the official random and generative benchmarks.

Table 2
Performance comparison between NixLab runs and official baselines (lower MAE is better).

Configuration / Run	MAE GAD-7	MAE PHQ-9	MAE CompACT-10	MAE Combined
NixLab Run 0 (Reactive)	0.800	1.180	1.593	1.191
NixLab Run 1 (Hybrid)	0.903	1.081	1.651	1.211
NixLab Run 2 (Historical)	1.014	1.005	1.711	1.243
BASELINE - Gemma	0.582	0.724	1.700	1.002
BASELINE - Random	1.185	1.215	2.123	1.508

All three variations of the architecture proposed outperformed the random baseline by a wide margin, validating the use of zero-shot NLI ensembles for this task. However, the internal trade-offs between the runs highlight a clear divergence in how specific clinical screening tools react to conversational context.

The Reactive configuration (Run 0), which placed 80% of its weight on the immediate user utterance, achieved our best overall performance with a combined MAE of 1.191. This run proved highly effective at mapping anxiety indicators, scoring an MAE of 0.800 on the GAD-7 scale. Anxiety symptoms in therapeutic dialogues are often triggered by acute verbal expressions and immediate panic markers, allowing the localized lexical booster and the responsive probability thresholds to capture these shifts instantly.

Conversely, as the influence of the historical dialogue window increased, performance on the anxiety and psychological flexibility scales (CompACT-10) degraded, while performance on the depression scale

(PHQ-9) improved significantly. The Historical configuration (Run 2) achieved the best PHQ-9 score with an MAE of 1.005, compared to 1.180 in the Reactive run. From a clinical perspective, tracking depressive states depends heavily on identifying sustained emotional baselines, persistent fatigue, or prolonged periods of low mood. Consequently, assigning a higher weight to the historical sliding window (45%) successfully smoothed out transient linguistic noise, preventing false positives and stabilizing predictions for chronic symptom indicators.

5.1. Sustainability and Efficiency Analysis

Operating zero-shot ensembles on moderate local hardware demands an evaluation of computational overhead. Table 3 presents the environmental metrics recorded by the background tracking layer during the evaluation sequence.

Table 3
Computational and environmental metrics recorded via CodeCarbon.

Metric Type	Measured Value
Total Wall Duration	2510 seconds
Energy Consumed	0.116 kWh
Carbon Emissions	0.0301 kg CO2-eq
Average CPU Energy	0.00371 kWh
Average GPU Energy	0.105 kWh

The total execution time across the evaluation rounds was 2510 seconds, drawing a total of 0.116 kWh of electrical energy. The localized GPU architecture bore the vast majority of the computational workload, consuming 0.105 kWh, whereas the CPU processing cores required a minimal draw of 0.00371 kWh. The resulting environmental impact was restricted to 0.0301 kg of CO2 equivalents. This footprint is remarkably low compared to larger generative models, demonstrating that targeted NLI ensembles provide an ecologically viable framework for real-time applications where cloud deployment is restricted due to patient privacy laws.

5.2. Error Analysis and Performance Limitations

An item-level breakdown of our performance reveals specific scenarios where our zero-shot architecture encounters systemic limitations. The most prominent failure mode is observed in the CompACT-10 instrument, where all three runs yielded significantly higher error rates (MAE between 1.593 and 1.711) compared to the GAD-7 and PHQ-9 scales.

We attribute this performance drop to two distinct factors:

1. **Scale Cardinality and Granularity:** CompACT-10 requires classification across a 7-point Likert scale (0 to 6). Because zero-shot NLI models yield continuous probability distributions that are inherently soft, partitioning the interval $[0, 1]$ into six distinct threshold boundaries (Θ_6) drastically narrows the margin of error for each class. A slight variation of 0.05 in the raw probability output often pushed predictions into an adjacent incorrect integer class, inflating the absolute error metric.
2. **Abstract Semantic Targets:** Unlike GAD-7 and PHQ-9 items—which map onto explicit somatic and emotional symptoms (e.g., *"insomnio"*, *"sentirse triste"*) that our lexical booster could directly capture—CompACT-10 targets highly abstract psychological mechanisms related to Acceptance and Commitment Therapy (e.g., avoidance strategies or mindfulness presence). Human dialogue rarely expresses concepts like "psychological flexibility" via simple explicit keyword roots. Consequently, our localized lexical booster had zero activation over these items, forcing the system to rely entirely on the base NLI model projections, which struggled to differentiate between subtle degrees of abstract clinical concepts in a zero-shot setting.

6. Conclusions and Future Work

In this paper, we presented a zero-shot multi-model ensemble system designed to translate unstructured patient dialogue into structured item-level psychometric data for the MentalRiskES task 1 at IberLEF 2026. By incorporating a localized history window, an adaptive trend-detection mechanism, and a targeted lexical booster, our system bypassed the need for domain-specific training data and significantly outperformed the random baseline across all evaluated clinical metrics.

The experimental configurations revealed a clear operational trade-off: short-term contextual windows coupled with aggressive thresholding yield optimal results for acute anxiety tracking (GAD-7), while long-term contextual integration is essential for capturing persistent depressive markers (PHQ-9). While our proposed localized models lagged behind large generative baselines like Gemma on certain sub-scales, our approach maintained a tiny carbon footprint of just 0.0301 kg of CO2 equivalents, emphasizing the long-term feasibility of smaller, dedicated architectures.

In future iterations of this framework, we plan to replace the hard step-threshold functions with continuous soft-voting calibration layers to better handle the fine-grained score boundaries of scales like CompACT-10. Additionally, we intend to explore the integration of dynamic context routing layers that automatically adjust history tracking weights based on the clinical dimensions detected at the start of the therapeutic session.

Declaration on Generative AI

During the preparation of this work, the author employed an advanced large language model (Gemini) strictly to assist in drafting, structuring, and formatting the LaTeX source document according to the required CEURART academic style guidelines. The author subsequently reviewed and edited the content as necessary and assume full responsibility for the accuracy and integrity of the published work.

References

[1] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2026: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.