Embedding and Heuristic-Based System for Psychometric Prediction and Response Selection in Therapeutic Dialogue

Ana Laura Lezama Sánchez[1,*,†], José Alejandro Reyes-Ortiz[2,†], Josué Padilla-Cuevas[2,†] and Gabriela Alejandra García-Robledo[2,†]

[1]*Secretaría de Ciencia, Humanidades, Tecnología e Innovacion, Av. Insurgentes Sur 1582, Col Credito Constructor, C.P.03940, Benito Juarez, Ciudad de Mexico, Mexico*

[2]*Universidad Autonoma Metropolitana, Unidad Azcapotzalco, Av. San Pablo Xalpa No. 180, Col. San Martín Xochinahuac, Alcaldia Azcapotzalco, C.P. 02128, Ciudad de Mexico, Mexico*

Abstract

In this paper we present a natural language processing approach for two therapeutic dialogue tasks proposed at IberLEF 2026. The proposed approach combines multilingual sentence embeddings with rule-based heuristic strategies to model semantic and contextual information from psychotherapy conversations. For Task 1, the system estimates PHQ-9, GAD-7, and CompACT-10 scores through cosine similarity between patient utterances and predefined clinical descriptors. For Task 2, the framework selects therapeutic responses by integrating contextual semantic similarity with clinically inspired heuristic rules related to empathy and conversational appropriateness. Based on the experimental results obtained on the official competition dataset, demonstrate stable performance across multiple evaluation rounds, achieving a combined MAE of 1.543 in psychometric prediction and an accuracy of 0.356 in therapeutic response selection. The results highlight the potential of lightweight embedding-based methods for supporting conversational mental health systems.

Keywords

Therapeutic dialogue, psychometric prediction, heuristic models, multilingual NLP,

CEUR-WS.org/Vol-4273/MentalRiskES2026_paper7.pdf

1. Introduction

The conversational Artificial Intelligence (AI) systems that help in the diagnosis of mental health have gained attention due to their potential to support psychological assessment and therapeutic interaction [1]. However, challenges remain in accurately interpreting patient language and generating appropriate responses that are both clinically meaningful and empathetic [2].

In recent years, the investigation on mental disease has increased considerably, but in the Spanish language, the dataset for Natural Language Processing (NLP) tasks is low, because datasets are more frequent in other languages [3].

Hence, this work addresses two core tasks proposed by IberLEF 2026 [4] [5], which consist of therapeutic dialogue systems: (1) psychometric prediction from natural language input and (2) response selection in multi-turn conversations. Therefore, we propose a hybrid approach combining pretrained multilingual sentence embeddings with heuristic rules to enhance both semantic understanding and interpretability. The main contributions of this work are: A multilingual embedding-based framework for automatic psychometric scale estimation from therapeutic conversations. A hybrid response selection strategy combining semantic similarity and therapeutic heuristics. An evaluation on the official IberLEF 2026 therapeutic dialogue tasks.

IberLEF 2026, septiembre de 2026, León, España

*Corresponding author.

†These authors contributed equally.

analaura.lezama@correo.buap.mx (A. L. Lezama Sánchez); jaro@azc.uam.mx (J. A. Reyes-Ortiz); jpc@azc.uam.mx (J. Padilla-Cuevas); gagr@azc.uam.mx (G. A. García-Robledo)

0000-0002-2505-5150 (A. L. Lezama Sánchez); 0000-0003-2111-4982 (J. A. Reyes-Ortiz); 0000-0002-3156-3231 (J. Padilla-Cuevas); 0000-0002-4559-651X (G. A. García-Robledo)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-09-29

The remainder of this paper is organized as follows. Section 2 presents the main concepts and theoretical background. Section 3 reviews related work in conversational AI and mental health applications. Section 4 describes the proposed methodology for both tasks. Experimental results are presented in Section 5, while Section 6 concludes the paper and discusses future directions.

2. Background

In this section, we present the main concepts that support our work, including conversational AI in mental health, natural language processing techniques, sentence embeddings, psychometric questionnaires, and multi-turn dialogue systems.

In recent years, conversational or textual AI systems are used increasingly more to support the evaluation of health, among them mental health and therapeutic interactions. These systems seek to understand the language of the patient and provide appropriate responses in a dialogue context. However, they face different challenges, such as capturing the semantic meaning of the expressions of the patient and generating coherent, empathetic, and clinically relevant responses.

To face these challenges, NLP plays a fundamental role by allowing machines to understand and process human language in a precise way. In the context of mental health, NLP techniques are applied to analyze clinical texts and therapeutic conversations. Such methods allow extracting meaningful information from unstructured data, which is essential for tasks such as symptom detection and response selection.

A key component of these approaches is the use of sentence embeddings, since they are vector representations of the text that capture the semantic meaning. Pretrained models, such as sentence transformers, allow the representation of sentences in a shared semantic space, which makes possible to calculate the similarity between the inputs of the patient and predefined elements using measures such as cosine similarity. Such representations are particularly useful in scenarios with limited resources, where labeled data are scarce.

In this context, psychometric instruments such as PHQ-9, GAD-7, and CompACT-10 provide standardized tools to evaluate mental health disorders, such as depression, anxiety, and psychological flexibility. Likewise, semantic similarity plays a fundamental role in the estimation of questionnaire responses, since it allows to use cosine similarity to compare the contributions of the patient with the items of the questionnaire.

3. Related Works

In this section, we review previous studies related to the application of natural language processing, machine learning, and artificial intelligence in mental health assessment, psychotherapy analysis, and therapeutic dialogue systems.

In [6] propose a NLP approach to automatically evaluate the quality of therapeutic sessions based on motivational interviews. The study uses 253 counseling transcripts classified into high- and low-quality sessions, from which 42 features related to conversational dynamics, semantic similarity, sentiment analysis, and question detection were extracted. Subsequently, the authors applied ML models such as Random Forest, CatBoost, and SVM, obtaining the best results with Random Forest, reaching up to 88.9% accuracy and 94.6% AUC after balancing and data augmentation techniques. The authors conclude that variables such as patient participation and semantic alignment between therapist and client are important indicators of therapeutic quality, demonstrating the potential of NLP and ML to support the objective evaluation of clinical conversations.

On the other hand, [7] present a review on the use of AI and ML in mental health, highlighting how these technologies are transforming traditional psychiatric diagnosis and treatment models. The authors point out that conventional evaluations depend largely on subjective assessments, which can generate misdiagnoses and delays in intervention. In response to this, the authors indicate that AI-based models use large volumes of data, behavioral analysis, and neuroimaging to offer more objective and predictive evaluations. In addition, they found evidence that techniques such as deep learning and NLP allow real-time monitoring, early disorder detection, and treatment personalization according to genetic, psychological, and environmental factors.

In [8] present a review of related works on the use of AI, ML, and deep learning for the detection, diagnosis, and treatment of depression. The authors highlight that data-based approaches allow the development of more precise and personalized systems in mental health, using information from text, magnetic resonance imaging, and functional magnetic resonance imaging. In addition, they managed to identify a transition from traditional ML models toward deep learning architectures and transformer-based embeddings, such as BERT, for clinical text analysis. Furthermore, the authors also indicate that AI can improve treatment response prediction and support therapeutic personalization; however, they warn about important limitations related to reduced sample sizes, overfitting, lack of external validation, and limited clinical evidence regarding real improvements in patients.

On the other hand, in [9] present that technologies such as NLP, deep learning, and multimodal data analysis from voice, text, and biosensors allow the identification of behavioral and physiological indicators associated with psychological distress. In addition, they propose concepts such as the digital psychological signature and digital mental health ecosystems to support more preventive and patient-centered care models. However, the authors indicate that the study also warns about limitations related to reduced sample sizes, lack of external validation, algorithmic biases, data privacy, and digital inequality.

In [10] present the use of ML and NLP as tools applied to the research and evaluation of psychotherapeutic processes. The authors analyze the capability of ML models to predict the therapeutic alliance from the linguistic content of psychotherapy sessions. For this purpose, they used recordings of 1,235 sessions corresponding to 386 patients treated by 40 therapists at a university counseling center, which were processed through automatic speech recognition. The authors mention that based on the obtained results, the algorithms managed to moderately predict therapeutic alliance ratings using only the verbal content of the sessions. In addition, the authors indicate that the study highlights the potential of AI to automate the evaluation of relevant clinical variables within psychotherapy and suggests that, with larger datasets and better technologies, these models could significantly support clinical training, therapeutic supervision, and the improvement of mental health services.

In [11] present a study that analyzes the use of ML algorithms to predict response to psychotherapeutic treatment in hospitalized patients within naturalistic clinical settings. The authors used a sample of 723 patients and evaluated demographic variables, physical indicators, psychological indicators, and treatment-related variables to identify factors associated with therapeutic success. Based on the obtained results, the authors showed a significant relationship between the initial severity of symptoms and the post-treatment condition, in addition to evidence that ML models outperformed traditional linear regressions in predictive capability. Among the most relevant factors, the authors highlighted treatment-related variables and psychological indicators, while physical and demographic variables had little influence.

The authors in [12] analyze the use of ML to predict patient responses to short-term dynamic therapies. The authors used data from 95 patients who participated in a 16-session treatment, monitoring progress weekly through the Outcome Questionnaire (OQ45) and Target Complaints (TC) instruments. Through random forest and elastic net models, the authors managed to identify change trajectories between responders and non-responders using pre-treatment information. Based on the obtained results, the authors showed that variables such as initial symptom severity, difficulties in emotion regulation, avoidant attachment, neuroticism, and interpersonal problems were important predictors in the OQ45 trajectories, reaching an accuracy of 72.8%. For the TC trajectories, the authors managed to highlight factors such as initial problem severity, self-sacrificing extraversion, and lack of assertiveness, with an accuracy of 62.8%.

4. Methodology

In this section, we present the methodology proposed for the two tasks of the MentalRiskES shared task at IberLEF 2026 [5, 4].

The first task, entitled: Early Symptom Detection in Therapeutic Conversations, is developed using the following methodology: For the development of task 1, an approach based on NLP was proposed to automatically estimate psychological indicators associated with the PHQ-9, GAD-7, and CompACT-10 scales from the therapeutic conversations provided by the competition. For the semantic representation of the text, a multilingual model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 was used. The model generates dense embeddings to capture semantic relationships between linguistic expressions in different languages.

The input data were the conversational transcripts stored in JSON format from the dataset provided by the competition, considering both the current patient message and the recent context of therapeutic interaction. To preserve the semantic continuity of the conversation, the last turns of the conversational history were integrated into the construction of the textual context used for the analysis. Subsequently, semantic embeddings were generated both from the patient's discourse and from the items corresponding to the evaluated clinical scales. The relationship between both vector spaces was calculated by applying cosine similarity, allowing the estimation of the degree of semantic correspondence between the conversational content and the psychological symptoms defined in each clinical instrument. The similarities were transformed into discrete scores through a heuristic function based on predefined thresholds. In the case of the CompACT-10 scale, score inversion was applied to the items defined as reversed according to the original structure of the instrument.

The second task: Therapist Response Selection (Decision Support) is developed with the next methodology: For the development of the second task consisted in the development of an automatic system for selecting therapeutic responses using semantic representations and heuristic strategies oriented to the clinical context. Therefore, the same multilingual transformer model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 from task 1 was used with the objective of maintaining consistency in the textual representation between both tasks. The system received as input the accumulated conversational history and the current patient message, which allowed the construction of a contextual representation of the therapeutic interaction. Subsequently, semantic embeddings were calculated for each one of the candidate responses provided by the system. The contextual relevance of each response was estimated through cosine similarity between the embedding of the conversational context and the embeddings of the available options. Additionally, a heuristic module based on linguistic patterns associated with desirable therapeutic practices was incorporated, including expressions of empathy, emotional validation, affective exploration, and open formulations oriented to clinical dialogue. In the same way, directive or potentially invalidating expressions were penalized.

Heuristic rules were incorporated to introduce clinically interpretable conversational behaviors that may not be fully captured through embedding similarity alone, particularly expressions related to empathy, emotional validation, and directive language. Figures 1 and 2 present the general frameworks of the proposed methodologies for both tasks.

5. Results

In this section we present the results obtained of applying the methodology proposed in section 4 for both tasks.

The proposed framework achieved stable performance across the three competition rounds. In Table 1 we expose the best PHQ-9 prediction performance was obtained in Round 1 with an MAE of 1.185, while the best GAD-7 performance was achieved in Round 2 with an MAE of 1.230. For CompACT-10, the lowest MAE was observed in Round 1 with a value of 2.173. Overall, the system maintained a consistent combined MAE of approximately 1.54 across all evaluations, demonstrating robustness and stability in automatic psychological scale estimation from therapeutic conversations.

Figure 1: General framework of the automatic clinical scale estimation task 1

Table 1
Performance results obtained for Task 1 across the three competition rounds.

Round	MAE GAD-7	MAE PHQ-9	MAE CompACT-10	Combined MAE
0	1.246	1.202	2.185	1.544
1	1.271	1.185	2.173	1.543
2	1.230	1.187	2.216	1.544

According to the official MentalRiskES 2026 leaderboard, our best submission (Run 1) achieved a Combined MAE of 1.543, obtaining the 31st position among participating systems. The best-performing system achieved a Combined MAE of 0.879. Although the proposed approach did not reach the top positions, the results demonstrate that a lightweight framework based on multilingual embeddings and heuristic rules can provide stable psychometric predictions without requiring extensive model fine-tuning.

The early detection evaluation assessed the capability of the proposed framework to estimate psychological scales at different conversational stages, represented by the R10, R30, and R60 evaluation windows. Table 3 we expose across the three competition rounds, the best overall early detection

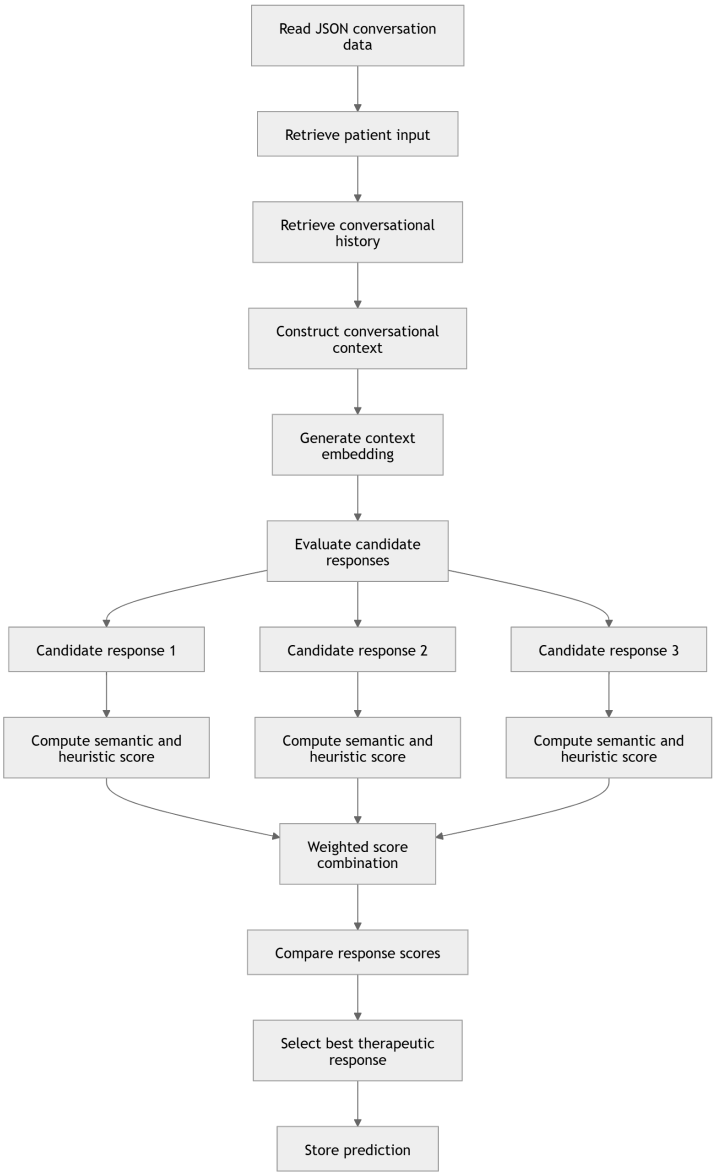

Figure 2: General framework of the automatic therapeutic response selection task 2

performance was achieved in Round 2 for the R10 setting with a combined MAE of 1.519, while the lowest combined MAE for the R60 setting was consistently obtained in Rounds 0 and 2 with a value of 1.157. On the other hand, the R30 setting produced the highest prediction errors across all evaluations, suggesting greater difficulty in estimating psychological indicators during intermediate conversational stages.

The evaluation of Task 2 (see Table 2) assessed the capability of the proposed framework to automatically select appropriate therapeutic responses during conversational interactions. Across the three competition rounds, the best predictive performance was achieved in Round 1, obtaining an overall accuracy of 0.356, a Macro F_1 of 0.353, and an Error Recovery Rate of 0.374. In contrast, round 2 produced the lowest overall performance with an accuracy of 0.303 and a Macro F_1 of 0.302. The results indicate that the combination of semantic similarity and heuristic therapeutic evaluation allowed the system to achieve stable response selection performance across different conversational settings.

The higher performance observed in Round 1 for Task 2 suggests that heuristic therapeutic patterns contributed positively to contextual response selection. However, the performance variability across rounds indicates that semantic similarity alone may not fully capture the complexity of therapeutic

Table 2
Performance results obtained for Task 2 across the three competition rounds.

Round	Accuracy	Macro Precision	Macro Recall	Macro F_1	Error Recovery Rate
0	0.319	0.319	0.321	0.319	0.315
1	0.356	0.353	0.355	0.353	0.374
2	0.303	0.305	0.302	0.302	0.289

dialogue dynamics.

In Table 3, we expose early detection performance results for Task 1 across the three competition rounds using MAE-Combined metrics at different conversational windows (R10, R30, and R60).

Table 3
Results for task 1 across the different rounds

Round	R10 MAE-Combined	R30 MAE-Combined	R60 MAE-Combined
0	1.603	2.440	1.157
1	1.547	2.040	1.200
2	1.519	2.090	1.157

5.1. Error Analysis

The results show that CompACT-10 consistently produced higher MAE values than PHQ-9 and GAD-7 across all evaluation rounds. This finding suggests that estimating psychological flexibility from conversational data may be more challenging than estimating depression or anxiety indicators. Furthermore, the variability observed across conversational windows indicates that the amount of available dialogue context can substantially influence prediction quality. Future work should investigate which linguistic and contextual factors contribute most strongly to these prediction errors.

6. Conclusions

In this work we presented an NLP approach for automatic psychological scale estimation and therapeutic response selection from psychotherapy conversations. The proposed approach combined transformer-based semantic embeddings with heuristic clinical strategies to model conversational context and psychological indicators. Experimental results obtained in the official competition demonstrated stable performance across multiple evaluation rounds, particularly for automatic estimation of PHQ-9, GAD-7, and CompACT-10 scales. Furthermore, the response selection task showed that integrating semantic similarity with therapeutic heuristics can support context-aware conversational decision-making. Although the obtained performance indicates remaining challenges in modeling complex therapeutic interactions, the proposed framework demonstrates the potential of NLP-based methods for supporting mental health assessment and conversational psychotherapy systems.

Due to the lightweight nature of the proposed framework, the heuristic component relied on manually defined conversational criteria rather than learned therapeutic strategies. Future work will explore more advanced contextual and dialogue-based models to improve both psychometric prediction and therapeutic response selection.

Declaration on Generative AI

The authors declare that no generative AI tools were used in the preparation of this manuscript.

References

[1] H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, D. C. Mohr, Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being, NPJ Digital Medicine 6 (2023) 236.

[2] H. Riess, G. Kraft-Todd, Empathy: a tool to enhance nonverbal communication between clinicians and their patients, Academic Medicine 89 (2014) 1108–1112.

[3] M. E. Villa-Pérez, L. A. Trejo, M. B. Moin, E. Stroulia, Extracting mental health indicators from english and spanish social media: a machine learning approach, IEEE Access 11 (2023) 128135–128152.

[4] 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.

[5] 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).

[6] A. Rueda, A. Perivolaris, N. Roy, D. Weston, S. Shaya, Z. Cote, M. Ivanov, B. G. Teferra, Y. Wu, S. Rambhatla, et al., Estimating quality in therapeutic conversations: A multi-dimensional natural language processing framework, arXiv preprint arXiv:2505.06151 (2025).

[7] S. Omiyefa, Artificial intelligence and machine learning in precision mental health diagnostics and predictive treatment models, Int J Res Publ Rev 6 (2025) 85–99.

[8] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, U. R. Acharya, Y. Li, Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment, Brain Informatics 10 (2023) 10.

[9] N. Mikaeili, M. Naeim, M. Narimani, Reimagining mental health with artificial intelligence: early detection, personalized care, and a preventive ecosystem, Journal of Multidisciplinary Healthcare (2025) 7355–7373.

[10] S. B. Goldberg, N. Flemotomos, V. R. Martinez, M. J. Tanana, P. B. Kuo, B. T. Pace, J. L. Villatte, P. G. Georgiou, J. Van Epps, Z. E. Imel, et al., Machine learning and natural language processing in psychotherapy research: Alliance as example use case., Journal of counseling psychology 67 (2020) 438.

[11] K. Jankowsky, L. Krakau, U. Schroeders, R. Zwerenz, M. E. Beutel, Predicting treatment response using machine learning: A registered report, British Journal of Clinical Psychology 63 (2024) 137–155.

[12] R. Yonatan-Leus, G. Gwertzman, O. Tishby, Using machine learning methods to identify trajectories of change and predict responders and non-responders to short-term dynamic therapy, Psychotherapy Research 35 (2025) 1070–1086.