RecEarlyMind @ eRisk 2026: Conversational Depression Detection Using Transformer-Based Language Models

Notebook for the eRisk Lab at CLEF 2026

Stergio Eugin[1], Beulah A[1], Sathvikha V[1] and Suhirtha M P[1]

[1]Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Chennai, India

Abstract

Depression is one of the most significant mental health challenges worldwide, making early detection an important research problem in natural language processing and artificial intelligence. This paper describes the participation of Team RecEarlyMind in CLEF 2026 eRisk Task 1 on Conversational Depression Detection. The task focuses on identifying signs of depression from conversational interactions generated through Large Language Model (LLM) personas. Our approach utilizes the Meta-Llama-3-8B-Instruct model together with conversational text analysis techniques to classify conversations into depression-related and non-depression-related categories. The proposed system includes preprocessing, contextual understanding, and prompt-based analysis for effective prediction. Experimental results demonstrate the effectiveness of our approach in detecting depressive symptoms from conversational data. This work highlights the potential of LLM-based systems for supporting early mental health risk assessment and conversational depression detection.

Keywords

Conversational Depression Detection, Large Language Models, Meta-Llama-3-8B-Instruct Model, eRisk 2026

1. Introduction

The CLEF 2026 eRisk Task 1 focuses on Conversational Depression Detection using Large Language Model (LLM) personas. The shared task aims to identify depressive symptoms from conversational interactions by analyzing contextual and emotional patterns present in dialogue sequences. The task and its evaluation framework are described in the official eRisk 2026 overview papers [1, 2]. Unlike traditional text classification tasks, conversational depression detection [3] requires systems to understand semantic relationships, emotional tone, behavioral indicators, and contextual dependencies within conversations. The task provides a realistic conversational setting in which participants must generate predictions while strictly following the official system prompts and evaluation guidelines defined by the organizers .

Our team, RecEarlyMind from Rajalakshmi Engineering College, participated in CLEF 2026 eRisk Task 1 by developing a conversational analysis framework based on the Meta-Llama-3-8B-Instruct model [4]. Our primary objective was to explore how Large Language Models and contextual conversational understanding techniques could be adapted for mental health-related classification tasks. The proposed system utilizes transformer architectures [5], prompt-based interaction strategies, and contextual feature extraction methods to identify depressive symptoms from conversation histories. The experiments were conducted using conversational datasets and LLM personas released by the organizers for weekly evaluation rounds.

Recent advancements in Natural Language Processing and transformer-based language models have significantly enhanced the ability of artificial intelligence systems to understand conversational context, semantic relationships, and emotional expressions. Transformer architectures have achieved state-of-the-art performance across a wide range of NLP tasks, including sentiment analysis, dialogue understanding, text classification, and mental health assessment. Their self-attention mechanism

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

231801171@rajalakshmi.edu.in (S. Eugin); beulah.a@rajalakshmi.edu.in (B. A); 231801160@rajalakshmi.edu.in (S. V); 231801175@rajalakshmi.edu.in (S. M. P)

0009-0005-0398-0790 (S. Eugin); 0000-0002-3891-0806 (B. A); 0009-0008-2220-3200 (S. V); 0009-0007-3531-1887 (S. M. P)

enables effective modeling of long-range contextual dependencies and nuanced linguistic patterns within conversations, making them well-suited for conversational depression detection. In addition, prompt engineering and contextual reasoning techniques further improve the capability of these models to extract psychologically relevant information from dialogue, supporting more reliable identification of depression-related symptoms.

Several studies have explored the use of NLP and deep learning techniques for mental health assessment from textual data, including approaches for depression detection [4]. Trotzek et al. investigated recurrent neural network architectures for depression detection from social media posts [6] and demonstrated the effectiveness of deep contextual representations for identifying depressive behavior. These studies emphasize the growing importance of AI-driven approaches in supporting early psychological risk assessment.

Research in conversational AI and mental health analysis also highlights the importance of contextual understanding, emotional reasoning, and ethical AI deployment. Turcan et. al., introduced benchmark datasets for mental health analysis and demonstrated how conversational context can improve depression detection performance [4]. Likewise, Zogan et. al., explored transformer-based architectures for mental health prediction and reported that contextual attention mechanisms improve the identification of depression-related linguistic patterns . These approaches indicate that transformer-driven conversational systems can effectively support automated mental health assessment tasks.

In this work, we present a conversational depression detection framework [7] based on transformer-driven contextual analysis techniques [5] and prompt-based conversational understanding. The proposed system includes conversational preprocessing, contextual feature extraction, emotional pattern analysis, and depression classification using Large Language Models [8]. The experiments were performed using the datasets and evaluation protocols provided in the CLEF 2026 eRisk Task 1 challenge. The obtained results demonstrate the effectiveness of transformer-based conversational analysis techniques for detecting depressive symptoms from dialogue interactions and highlight the potential of AI-assisted systems in supporting early mental health risk detection and conversational healthcare applications.

2. Experimental Setup

The experiments were conducted using transformer-based Large Language Models for conversational depression detection [4] in CLEF 2026 eRisk Task 1. The conversational data provided by the organizers was preprocessed by removing unnecessary symbols, whitespace characters, and irrelevant formatting information to improve textual consistency. The cleaned conversations were analyzed using contextual and prompt-based techniques to identify depressive symptoms from conversational interactions. The implementation was carried out using Python, Hugging Face Transformers, PyTorch, NumPy, Pandas, and Scikit-learn libraries. GPU-supported computation was used for efficient processing of conversational sequences and faster model inference. The official system prompts provided by the organizers were strictly followed without any modifications according to the task guidelines. The submitted system was evaluated using the official eRisk 2026 Task 1 evaluation metrics, namely DCHR, ADODL, ASHR, LDCHR, and LASHR. to measure the effectiveness of the conversational depression[9] detection system.

3. Approach

Our approach for CLEF 2026 eRisk Task 1 focuses on detecting depressive symptoms from conversational interactions using transformer-based Large Language Models and contextual conversational analysis techniques. The proposed system processes conversational data generated through predefined LLM personas and classifies conversations into *depression-related* and *non-depression-related* categories based on the linguistic, semantic, and emotional patterns identified throughout the dialogue history [10].

Initially, the conversational data undergoes preprocessing to improve textual consistency and model understanding. This stage includes text cleaning, removal of unnecessary symbols and special characters,

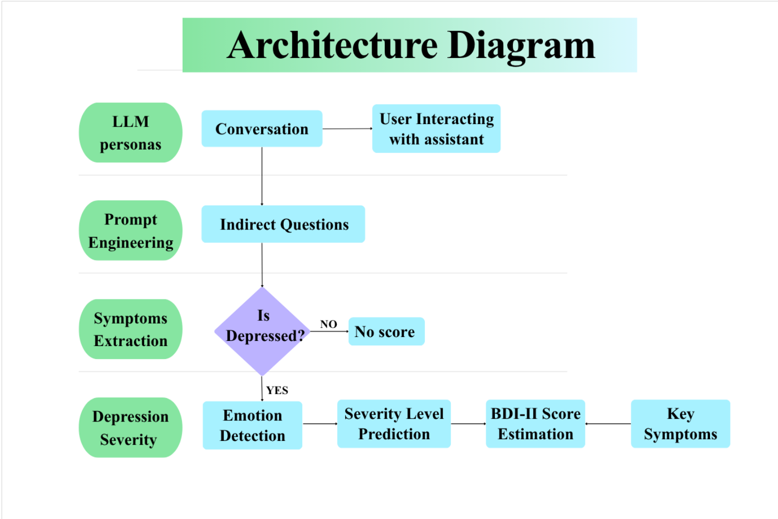

Figure 1: System Architecture for eRisk and Depression Detection

normalization, and tokenization of conversational sequences. The processed conversations are then converted into structured input representations suitable for transformer-based architectures. Contextual dependencies and semantic relationships across multiple conversation turns are preserved to improve conversational understanding.

The core component of the proposed framework employs the Meta-Llama-3-8B-Instruct model for conversational analysis and depression prediction. The model processes the complete conversation history to capture contextual dependencies, semantic relationships, emotional expressions, and psychologically relevant linguistic patterns associated with depressive behaviour. The official prompts provided by the eRisk 2026 organizers were used without modification to interact with the LLM personas throughout the evaluation. The contextual representations generated by Meta-Llama-3-8B-Instruct were then used to estimate the depression status, predict the BDI-II score, and identify the most relevant depression symptoms for each conversation.

The conversational context processed by the Meta-Llama-3-8B-Instruct model was used to identify depression-related linguistic and emotional patterns. Based on the complete conversation history, the model generated the depression prediction together with the estimated BDI-II score and the ranked depression symptoms. The predictions were produced according to the official output format specified for the eRisk 2026 Task 1.

The experiments were implemented using Python together with the Hugging Face Transformers library, PyTorch, NumPy, Pandas, and Scikit-learn. The official system prompts provided by the eRisk 2026 organizers were followed without modification, in accordance with the task guidelines. The submitted system was evaluated using the official eRisk 2026 Task 1 evaluation metrics, namely Depression Conversation Hit Rate (DCHR), Average Difference between Obtained and Desired Depression Level (ADODL), Average Symptom Hit Rate (ASHR), Latent Depression Conversation Hit Rate (LDCHR), and Latent Average Symptom Hit Rate (LASHR), which assess the quality of depression prediction and symptom identification.

3.1. Conversation Workflow

The submitted system employed the Meta-Llama-3-8B-Instruct model as the primary conversational language model for eRisk 2026 Task 1. During the evaluation, the system interacted directly with the LLM personas provided by the organizers for each weekly evaluation round using the official interaction protocol and prompts without any modifications. The complete conversation history was preserved as context for every interaction, allowing the model to capture semantic relationships, emotional expressions, and contextual dependencies across dialogue turns.

The submitted system did not implement an early stopping strategy. Conversations continued until the organizer-defined interaction process was completed, after which the final conversation history was used for prediction. Based on the complete dialogue history, the Meta-Llama-3-8B-Instruct model generated the depression prediction, estimated the BDI-II score, and ranked the depression symptoms according to the official task prompt and required output format.

4. Results and Analysis

The performance of the proposed conversational depression detection [1] system was evaluated using the official evaluation framework provided in CLEF 2026 eRisk Task 1. The task focused on identifying depressive symptoms from conversational interactions with LLM personas while encouraging early decision making [11]. The evaluation considered multiple metrics including Depression Case Hit Rate (DCHR) [9] (DCHR), Average Depression Ordinal Distance Loss (ADODL), Average Symptoms Hit Rate (ASHR), Latency-weighted Depression Case Hit Rate (LDCHR), and Latency-weighted Average Symptoms Hit Rate (LASHR). These metrics measure not only the correctness of predictions but also the efficiency of conversational interactions before generating predictions.

4.1. Results for eRisk: Conversational Depression Detection

Our team, **RecEarlyMind**, participated with one official run in the shared task evaluation. The obtained results are presented in Table 1. The proposed approach achieved a DCHR score of 0.3000, an ADODL score of 0.8071, an ASHR score of 0.1625, an LDCHR score of 0.2013, and a LASHR score of 0.0977. The results demonstrate the capability of the proposed Meta-Llama-3-8B-Instruct-based conversational analysis framework in identifying depressive symptoms from dialogue interactions generated by LLM personas.

Table 1
Evaluation Results for RecEarlyMind in CLEF 2026 eRisk Task 1

Team	DCHR	ADODL	ASHR	LDCHR	LASHR
RecEarlyMind	0.3000	0.8071	0.1625	0.2013	0.0977

Table 2
Participation Statistics of Team RecEarlyMind

Team	**Runs**	**Personas/run**	**Manual**	**Msgs./Conv.**	**Chars./Msg.**
RecEarlyMind	1	20	–	7.5	93.23

The experimental findings indicate that contextual understanding and transformer-based conversational analysis can effectively support depression detection from conversational data. However, the results also highlight the challenges associated with accurately identifying depressive symptoms within limited conversational interactions. Future improvements may include enhanced prompt engineering strategies, larger conversational context modeling, and advanced emotional reasoning techniques for improving prediction accuracy and early risk detection performance[11].

5. Example Outputs

Table 3 illustrates representative examples of the system predictions for conversational inputs. The examples are provided to demonstrate the format of the model outputs and the types of depression-related indicators identified by the proposed system.

Table 3
Example Outputs of the Proposed System

Conversation Snippet	Predicted Class	Detected Indicators
"I feel tired all the time and I do not enjoy things anymore."	Depression	Hopelessness, Loss of Interest
"I have been avoiding my friends recently and prefer staying alone."	Depression	Social Withdrawal, Isolation
"I am doing well and spending time with my family and friends."	Non-Depression	Positive Emotional State
"Sometimes I feel worthless and cannot focus on my daily activities."	Depression	Negative Self-Perception, Low Motivation
"I completed my work today and felt productive."	Non-Depression	Positive Behaviour

The generated outputs demonstrate that the proposed system can identify contextual emotional patterns and conversational cues associated with depressive behaviour. The proposed Meta-Llama-3-8B-Instruct framework effectively distinguishes depression-related and non-depression-related interactions using contextual semantic understanding and conversational reasoning.

Table 4
Example Outputs of Conversational Depression Detection

Conversation Summary	BDI Score	Detected Symptoms
The user expressed tiredness, sleep disturbances, work stress, difficulty relaxing, and constant overthinking during the conversation.	12	Sadness, Indecisiveness, Irritability, Concentration Difficulty
The user reported social withdrawal, lack of motivation, feelings of worthlessness, and difficulty concentrating on daily activities.	18	Social Withdrawal, Low Self-Esteem, Sadness, Sleep Disturbance

6. Conclusion

In this paper, Team RecEarlyMind presented a transformer-based conversational depression[9] detection framework for CLEF 2026 eRisk Task 1. The proposed approach utilized Large Language Models, contextual conversational analysis, and prompt-based techniques to identify depressive symptoms from dialogue interactions generated through LLM personas. The system focused on analyzing emotional expressions, contextual dependencies, and linguistic patterns associated with depression[3] to classify conversations into depression-related and non-depression-related categories.The experimental results demonstrate that transformer-based[5] conversational analysis techniques can effectively support early depression[9] detection from conversational data. The obtained evaluation scores highlight the potential of contextual language understanding models for mental health analysis and conversational AI applications. Although the task presents several challenges due to limited conversational interactions and complex emotional patterns, the proposed framework achieved meaningful performance in identifying depressive indicators.Future work may focus on improving contextual reasoning, advanced prompt engineering strategies, emotion-aware conversational modeling, and larger transformer architectures to further enhance prediction accuracy and early risk detection capabilities[11].

Acknowledgements

The authors would like to express their sincere gratitude to the organizers of the CLEF 2026 eRisk Task 1 for providing the datasets, evaluation framework, and research platform for conversational depression detection [3]. We also thank Rajalakshmi Engineering College, Department of Artificial Intelligence and Data Science, for providing the necessary support and resources to carry out this research work. Finally, we acknowledge the open-source NLP and deep learning communities, including Hugging Face and PyTorch, for providing tools and frameworks that supported the implementation of this work.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to improve grammar, sentence structure, language clarity, and content refinement. The authors also used Generative AI tools for assistance in drafting and organizing portions of the manuscript. After using these tools, the authors carefully reviewed, edited, and validated the content as needed and take full responsibility for the final publication content.

References

[1] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[3] F. Sheikhi, L. Fakher, D. Chekani, Depression detection on e-risk 2017 using long short-term memory models, in: 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), IEEE, 2024, pp. 1–6.

[4] M. Saad, M. Abbas, A. Chaudhry, F. Alvi, A. Samad, Contextualized early detection of depression–hybrid and time-aware approaches: Hu at erisk task 2 2025, Working Notes of CLEF (2025) 9–12.

[5] S. Adhikary, J. Das, D. Roy, Thinkir at erisk 2025: early detection and risk assessment of depression using transformer models, Working Notes of CLEF (2025) 9–12.

[6] M. Stankevich, V. Isakov, D. Devyatkin, I. V. Smirnov, Feature engineering for depression detection in social media., in: ICPRAM, 2018, pp. 426–431.

[7] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020 (2020).

[8] J. Parapar, A. Perez, X. Wang, F. Crestani, erisk 2025: contextual and conversational approaches for depression challenges, in: European Conference on Information Retrieval, Springer, 2025, pp. 416–424.

[9] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2024: Depression, anorexia, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2024, pp. 474–481.

[10] E. Kara, R. E. M. Peña, L. Raithel, Fu-tu-dfki@ erisk 2025: a linguistically informed but overdiagnosing approach to early depression detection, Working Notes of CLEF (2025) 9–12.

[11] H. Damm, M. Trotzek, S. Koitka, H. Schäfer, A. Idrissi-Yaghir, C. M. Friedrich, erisk early risk detection of depression, Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project 1018 (2022) 61.