Early Depression Detection from Contextualized Online Conversations using Sequential Modeling*

Notebook for the eRisk Lab at CLEF 2026

Eman Sedqy Shlkamy[1,*], Wajdi Zaghouani[1]

[1]Northwestern University Qatar, Doha, Qatar

Abstract

Early detection of depression in online conversations presents a unique challenge that requires balancing prediction accuracy with timeliness of intervention. Unlike traditional text classification,early risk detection systems must operate under incomplete and evolving user histories,making sequential modeling essential. In this paper, we present our participation in eRisk 2026 Task 2, addressing depression detection from contextualized conversational data through a sequential modeling framework. Our system leverages TF-IDF representations combined with Logistic Regression and an evidence accumulation strategy that enables timely and reliable early alerts. On the training set, our model achieves an AUC-ROC of 0.944 and a macro F1-score of 0.78, demonstrating strong discriminative performance despite the inherent class imbalance and sequential constraints of the task.These results highlight that carefully designed lexical features,when paired with an effective decision strategy, can serve as a robust foundation for early risk detection in real-world conversational settings

Keywords

Natural Language Processing, Computational Social Science, TF-IDF, Logistic Regression, Early Depression Detection, eRisk, Sequential Modeling,

1. Introduction

By the time a person is diagnosed with depression, they may have been silently signaling distress for months through the words they write, the communities they engage with, and the conversations they hold online.Early risk detection systems aim to close this gap, identifying at-risk individuals before crisis points are reached.Early identification of depressive symptoms is critical for enabling timely intervention and improving long-term outcomes. With the widespread use of social media platforms,users increasingly share personal experiences, emotions and struggles online, providing valuable behavioral and linguistic signals that can support automated mental health assessment. Traditional approaches to depression detection typically frame the problem as a static text classification task,where a model is trained on a fixed collection of user-generated texts. However,real-world scenarios require systems that can detect risk as early as possible, based on partial and incrementally available information.This introduces the challenge of early risk detection, where models must balance predictive accuracy with timeliness of intervention. The eRisk 2026 Task 2 [1, 2] focuses on the early detection of depression from full conversational contexts.Unlike previous setups that rely solely on user-generated content, this task incorporates the entire conversational thread,requiring systems to model both the target user's writings and the surrounding interactions. This setting more closely reflects real-world monitoring scenarios,where user behavior is embedded within broader social exchanges.Early detection in such settings introduces several challenges. First systems must operate under incomplete and evolving data, where signals of depression may appear gradually over time. Second conversational context can introduce both noise and useful cues, making it difficult to distinguish relevant patterns.Third,early decision-making requires careful trade-offs, as premature predictions may reduce reliability, while delayed alerts limit practical usefulness.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author
eman.shlkamy@northwestern.edu (E. S. Shlkamy); wajdi.zaghouani@northwestern.edu (W. Zaghouani)
0009-0009-4542-3991 (E. S. Shlkamy); 0000-0003-1521-5568 (W. Zaghouani)

Research Gap While prior work has demonstrated strong performance using lexical and contextual representations for depression detection, most approaches assume access to complete user histories and do not explicitly model sequential decision-making. Moreover, the role of conversational context in early detection remains underexplored,particularly in settings where predictions must be made incrementally. Our work addresses this gap by proposing a sequential framework that operates on evolving conversational contexts

Our Contribution In this paper,we present a sequential modeling framework for early depression detection that leverages TF-IDF representations and Logistic Regression with an evidence accumulation strategy. Our system achieves an AUC-ROC of 0.944 on the training set, demonstrating that carefully engineered lexical features,combined with an effective decision strategy, yield strong and interpretable results under the sequential constraints of real-world early detection. Our contributions can be summarized as follows:

- We propose a sequential modeling framework for early depression detection that achieves strong discriminative performance through TF-IDF representations and Logistic Regression.
- We introduce an evidence accumulation decision strategy that enables timely and reliable early alerts across sequential rounds.
- We provide an analysis of model behavior under temporal constraints, offering insights into the trade-offs between detection speed and prediction reliability.

In the most recent edition of this task (eRisk 2025),top-performing systems relied on transformer-based architectures and risk signal fusion strategies,achieving F1 scores up to 0.85 [3] . Our work demonstrates that a carefully engineered lexical framework can deliver competitive performance under the same sequential constraints, with significantly lower computational cost.

2. Task Overview

The eRisk 2026 shared task on Contextualized Early Detection of Depression focuses on identifying signs of depression from sequential conversational social media data. Unlike traditional approaches that analyze isolated posts,this task emphasizes contextualized interactions involving multiple speakers and evolving conversational dynamics. Participants are required to process conversations incrementally, where the meaning and relevance of a message may depend on the surrounding conversational context. The objective is to detect depressive signals as early as possible while maintaining reliable classification performance. The task evaluates systems under an early risk detection framework, encouraging models to make accurate predictions using limited contextual evidence. This setting reflects real-world scenarios where early intervention is critical for mental health support. In this work,we propose a sequential framework for early depression detection in online conversations, combining TF-IDF lexical representations with a Logistic Regression classifier and a temporal evidence accumulation strategy for early decision-making. Our system is designed to balance prediction accuracy with detection timeliness, which represents one of the main challenges in early mental health monitoring tasks. Despite its simplicity, our approach demonstrates that interpretable lexical models remain highly competitive in temporally evolving conversational environments. The proposed framework provides a lightweight and effective baseline for contextualized early depression detection.

3. Related Work

Early risk detection on the internet has been extensively studied within the eRisk shared task series [4].

Prior work has explored a range of approaches,from lexical and rule-based methods to deep learning models,for detecting signs of depression from user-generated content on social media platforms. Lexical and feature-based approaches have demonstrated strong performance in early detection settings. Liu et al.[5] conducted a systematic review of machine learning methods applied to social media text for

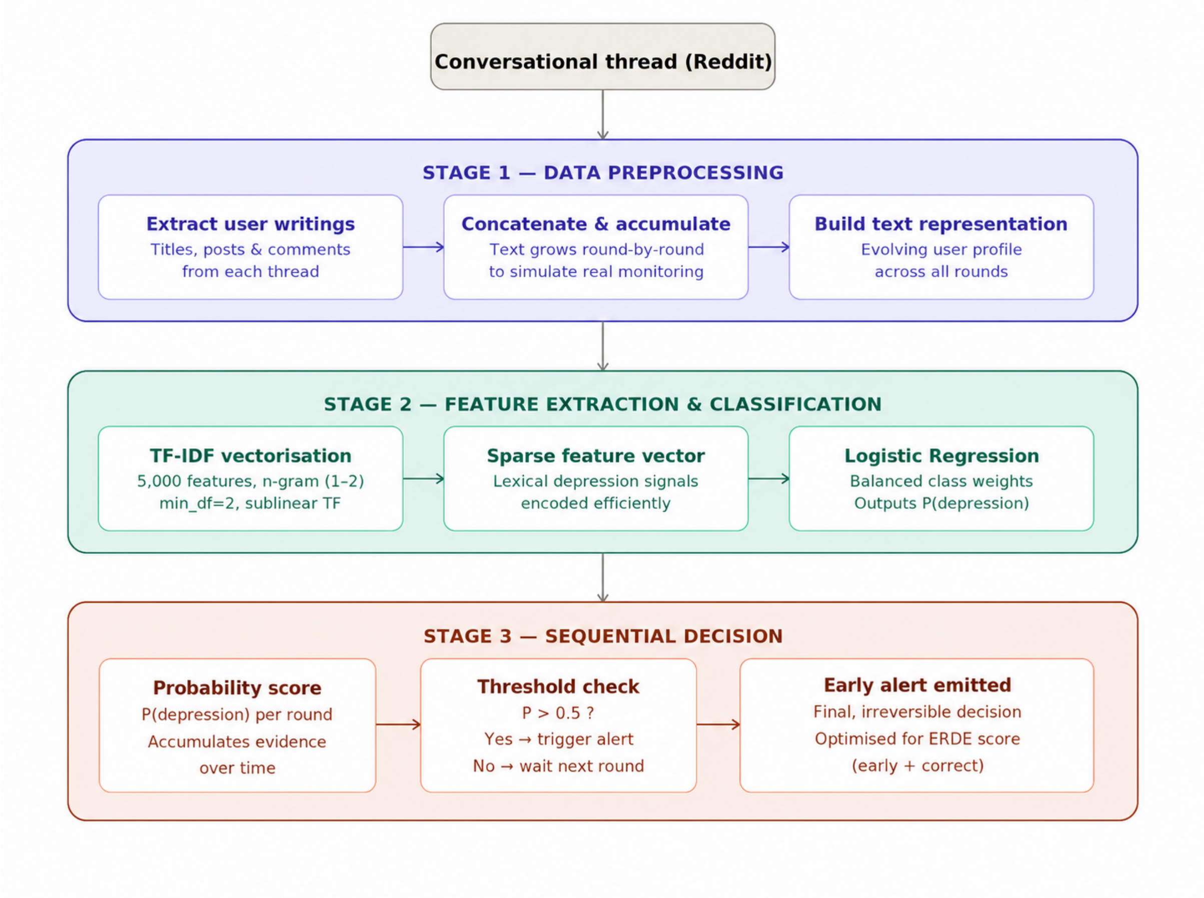

Figure 1: System pipeline for early depression detection in eRisk 2026 Task 2

depression detection,highlighting the effectiveness of NLP-based approaches as complementary tools in public mental health practice.Similarly,et al. [6] evaluated multiple classifiers including Logistic Regression, Naive Bayes, DistilBERT, and RoBERTa on a large-scale Twitter dataset of 632,000 tweets, demonstrating that traditional machine learning methods remain competitive despite the availability of transformer-based models.

Transformer-based architectures have further advanced the field.Martínez-Castaño et al. [7] proposed BERT-based models for early detection of mental health illnesses within the eRisk framework,achieving strong performance while highlighting the importance of sequential decision-making. Thekkekara et al. [8] introduced an attention-based CNN-BiLSTM model for depression detection on social media text,achieving competitive results by combining convolutional and recurrent architectures.Alsaedi and Yafooz [9] provided a comprehensive review of early depression detection approaches,highlighting the superiority of transformer-based models in capturing context-rich linguistic markers of depression.

In the most recent edition of this task (eRisk 2025), top-performing systems relied on transformer-based architectures and risk signal fusion strategies,achieving F1 scores up to 0.85 [3]. Despite these advances, most existing approaches assume access to complete user histories and do not explicitly model the sequential and incremental nature of early risk detection. Our work addresses this gap by proposing a sequential framework that operates on evolving conversational contexts with significantly lower computational overhead.

4. System Description

Figure 1 illustrates the three stage architecture of our system. Our system consists of three main components: data preprocessing, feature extraction, and sequential decision-making. This work emphasizes that early detection is not only about accuracy, but about making reliable predictions early enough to enable meaningful intervention. The design prioritizes interpretability and temporal awareness,processing user writings incrementally across rounds to simulate real-world monitoring conditions.

4.1. Data Preprocessing

For each target subject, we extract only the writings authored by the target user from each conversational thread. Specifically,we collect the body and title of submissions as well as all comments written by the user. The extracted texts are concatenated and accumulated across rounds to construct an evolving textual representation. This design enables the model to simulate real-world monitoring scenarios,where information becomes available incrementally over time.

4.2. Feature Extraction

We represent the accumulated text using TF-IDF vectorization with a maximum vocabulary size of 5,000 features, including unigrams and bigrams (n-gram range 1–2), a minimum document frequency of 2,and sublinear TF scaling. This configuration captures both single-word depression markers and multi-word expressions that frequently characterize depressive discourse, while sublinear scaling reduces the dominance of highly frequent terms. TF-IDF further supports efficient incremental updates as new conversational data becomes available across rounds, making it well-suited for the sequential nature of the task.

4.3. Classification

We train a Logistic Regression classifier with balanced class weights to address the significant class imbalance in the dataset (102 depressed vs. 807 control users). Logistic Regression was selected for its ability to produce well-calibrated probability scores, which are essential for the evidence accumulation strategy described in Section 4.4. The model outputs $P(\text{depression})$ at each round, enabling continuous risk monitoring over time.

4.4. Decision Strategy

At each round, if the predicted probability exceeds a threshold of 0.5, the system emits an alert. Once triggered, the decision remains final, following the task protocol. A decision threshold of 0.5 was applied, selected based on validation performance on the training set, where it yielded the best balance between precision and recall for the depression class. This strategy balances early detection with prediction confidence, aiming to reduce both delayed predictions and false positives.

5. Experimental Setup

5.1. Dataset

We use the eRisk 2025 contextualized dataset, consisting of 909 users (102 depressed and 807 control), reflecting a natural class imbalance ratio of approximately 1:8. Each user is represented as a sequence of conversational threads ordered chronologically. For each thread, we extract only the writings authored by the target user, including submission titles, bodies and comments.

5.2. Training Setup

We split the dataset into 80% training and 20% testing using stratified sampling to preserve the class distribution across splits. The TF-IDF vectorizer and classifier are trained on the training split and evaluated on the held-out set.

5.3. Evaluation Metrics

We report AUC-ROC, macro F1-score, precision, and recall. In addition, the official evaluation uses the ERDE metric,which penalizes both incorrect and late predictions. The ERDE metric reflects real-world constraints, where timely detection is as important as accuracy.

6. Results and Analysis

6.1. Training Set Performance

Table 1 presents the performance of our model on the held-out validation set.The model achieved an AUC-ROC of 0.944 , demonstrating strong discriminative capability between depressed and control users despite the significant class imbalance of the dataset.

Table 1
Performance on the held-out training set.

Class	Precision	Recall	F1	Support
Control	0.96	0.93	0.95	162
Depression	0.56	0.70	0.62	20
Macro Avg	0.76	0.82	0.78	182

For the control class, the model achieved high precision (0.96) and recall (0.93), yielding an F1-score of 0.95. For the depression class,recall reached 0.70,reflecting a meaningful ability to identify at-risk users.The lower precision (0.56) is expected given the 1:8 class imbalance and the inherent overlap between depressive and non-depressive linguistic patterns. The macro F1 of 0.78 confirms balanced performance across both classes. We observe that misclassifications are often associated with context ambiguity,where conversational replies from other participants introduce noise that may obscure the target user's true mental state.

7. Discussion

The strong AUC-ROC of 0.944 suggests that depression-related language is characterized by consistent and detectable lexical patterns,including expressions of negative affect, hopelessness and self-referential language. These signals remain identifiable even under the sequential constraints of the task

The lower precision for the depression class highlights a key limitation: lexical features alone cannot fully capture the complexity of mental health expressions in conversational settings.Contextual and temporal signals are likely necessary to reduce false positives. Moreover, early detection performance must be evaluated beyond classification accuracy the timing of predictions is equally critical, as delayed alerts reduce practical usefulness while premature alerts risk false alarms. Note that all reported results are evaluated on a held-out portion of the training set. Official test set results are pending the eRisk 2026 task evaluation.

8. Conclusion

This paper presented a sequential framework for early depression detection in contextualized online conversations, developed as part of eRisk 2026 Task 2 . Our system combines TF-IDF lexical representations with Logistic Regression and an evidence accumulation strategy that enables timely,round-by-round risk assessment. The system achieved an AUC-ROC of 0.944 and a macro F1-score of 0.78 on the held-out training set,establishing that carefully engineered lexical features when paired with an effective sequential decision strategy yield strong and interpretable results under real-world monitoring constraints. Our analysis reveals a key trade-off inherent to early detection: optimizing for recall risks increasing false positives, while conservative thresholds delay intervention.Addressing this trade-off remains an open challenge that future work should tackle through adaptive thresholds, contextual embeddings,and richer behavioral features. Ultimately, this work demonstrates that interpretable and efficient systems can serve as strong foundations for computational mental health monitoring and that advancing the field requires moving toward temporally aware, context-sensitive architectures.

9. Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF) part of Qatar Research Development and Innovation Council (QRDI)

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT as a writing assistant in order to Grammar and spelling check. Further, the author(s) used claude for figures 1 in order to: Generate images.After using these tool(s)/service(s),the author(s) reviewed and edited the content as needed and take(s) 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: Symptom ranking and conversational approaches for depression and adhd (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, 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] T.-P. Mai, M.-H. H. Le, D.-L. Tran, D.-C. Can, H.-Q. Le, Uet@ erisk2025: Severity estimation for depression symptoms searching and early risk detection, Working Notes of CLEF (2025) 9–12.

[4] D. E. Losada, F. Crestani, J. Parapar, erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 346–360.

[5] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, J. Guo, et al., Detecting and measuring depression on social media using a machine learning approach: systematic review, JMIR Mental Health 9 (2022) e27244.

[6] B. G. Bokolo, Q. Liu, Deep learning-based depression detection from social media: comparative evaluation of ml and transformer techniques, Electronics 12 (2023) 4396.

[7] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Bert-based transformers for early detection of mental health illnesses, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 189–200.

[8] J. P. Thekkekara, S. Yongchareon, V. Liesaputra, An attention-based cnn-bilstm model for depression detection on social media text, Expert systems with applications 249 (2024) 123834.

[9] A. Alsaedi, W. M. S. Yafooz, Early depression detection from social media: State-of-the-art approaches, in: AI-Driven: Social Media Analytics and Cybersecurity, volume 1180 of *Studies in Computational Intelligence*, Springer, 2025, pp. 61–75. doi:10.1007/978-3-031-80334-5_4.