SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Temporal Language Modelling For Early Depression Risk Detection

Nisha F[1]*, Radhakrishnan Vignesh[2]

[1,2] Presidency School of AI and Advanced Computing, Presidency University Itgalpur, Rajankunte, Yelahanka, Bengaluru – 560119, Karnataka, India Email: vigneshr@presidencyuniversity.in

*Corresponding Author: Nisha F Email:NISHA.20233CSE0017@presidencyuniversity.in;
nisharobinr@presidencyuniversity.in
ORCID: https://orcid.org/0000-0001-9538-3048

Abstract:
This requires the development of models that go beyond snapshot-based classifications to account for the temporal evolution of language use. In this paper, we propose a temporal trajectory learning pipeline using contextual language embeddings to represent the sequential evolution of user posts, which uses a GRU-based recurrent temporal layer to learn the evolution of depression risk over time. We use a patience-based decision policy to make early risk decisions, which are typically based on separate thresholds. We use various metrics, including time-to-detection statistics, the proposed ERDE metric in compliance with the benchmark, and the usual classification metrics. We test the proposed temporal trajectory learning pipeline using the official CLEF 2017 eRisk benchmark as the primary evaluation corpus, as well as the Dreaddit corpus as a secondary generalization test. We perform an ablation analysis using three different baseline models: TF-IDF + Logistic Regression, LSTM-only, and BERT-Final-only. We find that the proposed model performs best on the eRisk 2017 benchmark, obtaining AUC = 0.8852, ERDE_50 = 0.0775 in compliance with the benchmark and outperforms the LSTM-only variant by 0.37 AUC and the BERT-Final-only variant by 0.14 AUC.

Keywords: depression risk detection, temporal modelling, early risk detection, social media analysis, GRU networks, benchmark validation.

1. INTRODUCTION

Depression is a chronic and often episodic illness in which risk factors may develop gradually through changes in emotions, cognition, and behavior. Social media platforms retain a history of user expression and activity that can be mined to forecast risk factor trends. However, existing techniques for mental health detection are mostly static classification models that attempt to classify a given text snippet into a single class. This is a simplistic model that does not account for the progressive nature of depression indicators, where a patient's condition
worsens over a period, resulting in a change in sentiment, language, and activity patterns. Temporal trajectory modeling of depression indicators redefines depression detection as a sequence of predictions. Instead of a single classification, a model can be used to generate a series of predictions that indicate a patient's risk factor over a period. This allows for early warnings and estimation of lead times. Lead times are valuable in real-world depression detection, where early detection is more valuable than a slight increase in classification accuracy at a single point in time. Therefore, this research aims to build a temporal pipeline that provides a series of predictions, allows for early decision, and provides interpretable risk factor trajectories. The proposed temporal pipeline is validated on two datasets. One is the official CLEF eRisk 2017 benchmark for early depression detection, where user timelines are provided in ten segments of equal time duration, thus allowing for ERDE evaluation. As a secondary dataset, a Reddit-derived corpus named "Dreaddit" is used to evaluate pipeline behavior for large data conditions. An ablation study on three variants of a baseline model is conducted to evaluate the performance of the DistilBERT encoder and GRU

temporal layer. We hypothesize that temporal trajectory modelling combined with patience-based early decision policies can reduce false positives and improve early depression risk detection compared to snapshot classification approaches.

The existing depression detection approaches on social media primarily take the form of snapshot text classification. This paradigm fails to capture the changing behavioral and linguistic patterns effectively and fails to provide defensible early risk lead-time estimates. There exists a need for a reproducible temporal framework that provides stepwise risk estimates, facilitates early decision-making policies, and follows the early risk benchmarks.

Create a reproducible temporal trajectory pipeline that produces stepwise risk prediction. Combine contextual linguistic representations with a temporal sequence model to make user-level risk predictions. Incorporate early decision mechanisms and provide performance and time-to-detection metrics. Validate the overall pipeline with the official eRisk 2017 benchmark dataset by utilizing the ERDE evaluation tool, followed by secondary validation on the Dreaddit dataset with an ablation study to understand the component-wise performance and the influence of different conditions on the overall performance.

2. LITERATURE SURVEY

Longitudinal modeling of changes in the depressive state has also seen increased prominence in the context of its ability to capture changing risks rather than fixed categories. Recent studies have highlighted the importance of monitoring longitudinal changes through language and
sequential signal processing, including speech and textual features [1]. Models that specifically address the semantics of time series have also seen increased sensitivity in the context of longitudinal symptom changes, where the structural aspect of time has seen a positive impact [2]. Hybrid approaches that combine behavioral and linguistic features have also seen a positive development in the context of depression prediction, specifically in the context of social media where activity patterns also indicate meaningful changes [12], [14]. Explainability has also seen increased relevance in the context of mental health prediction, where temporal networks have seen the development of explainable reasoning and user-level temporal evidence [3]. Furthermore, multimodal learning frameworks also increasingly seek to address the fusion of context-driven representations [3], [7]. Apart from the Reddit platform, the Weibo platform has also seen the development of depression prediction pipelines, highlighting the differences that may exist in the context of different social media platforms [5]. Recent studies have also highlighted the development of personalization in the context of mental health prediction for adolescents, highlighting the positive development in the context of mental health prediction [6]. Large language model frameworks also increasingly seek to address mental health prediction tasks, including the development of multimodal and multi-task approaches [7], [19] and the development of longitudinal perspectives in the context of mental health prediction [1], [2]. Embedding-based multimodal approaches also highlight the positive development in the context of mental health prediction [13], [22].

Early risk detection benchmarks have been at the core of influencing evaluation approaches for depression detection systems. The eRisk initiative, conducted under the CLEF evaluation forum, has offered standardized early risk prediction tasks since 2017, defining Early Risk Detection Error (ERDE) as a main evaluation metric for temporally informed systems [8], [11]. As opposed to traditional classification benchmarks, the eRisk evaluation framework demands depression detection systems to process user writing incrementally, producing risk judgments as early as possible. Thus, it explicitly penalizes delay in detecting depression by incorporating ERDE in its evaluation formula. This evaluation framework has identified a significant disparity in the performance of classification systems that optimize for end accuracy and early risk detection systems that must strike a delicate balance between end accuracy and early detection. Progressing through the evaluation framework of the eRisk task in consecutive conferences has identified that a large proportion of participating systems are closer to classification systems than actual trajectory learning systems. Transformer models have also been explored for mental health detection problems [18]. Language models like BERT and its variants have been found to perform well for depression classification problems [9], [15].

Context-driven multimodal frameworks have also been introduced for depression detection [4]. Their work shows that richer context representations can lead to better accuracy in depression detection. Zafar et al.
[3] introduced Multi-Explainable TemporalNet for depression detection. Their model leverages multimodal inputs and explainability using attention mechanisms. Although this model leverages temporal information like other models, it is quite different from the model introduced here because it does not allow stepwise probability calculations for decision-making. Moreover, it does not allow a patience mechanism that can help in early decision-making with a detectable delay.

Cross-platform and cross-population generalization is an issue that has yet to be fully addressed in the field of social media depression detection. Zhang et al. [5] have shown that a depression prediction pipeline using

NLP may not automatically be portable between platforms such as Reddit and the Chinese microblogging platform Weibo. This is because of the inherent differences in language usage and posting patterns between the two platforms. This is in line with the current study's use of dual-dataset validation on eRisk 2017 and Dreaddit datasets. Zhang and Li [6] further emphasized the importance of personalization in the field of risk prediction for the adolescent demographic, in whom the differences between individual and population-level behavior are the greatest. All of the above studies support the development of modular temporal prediction pipelines that can be adapted for different populations without the need for changes in the overall structure of the model. Recurrent architecture in the field of sequential risk prediction have been well established in the eRisk community. In the early stages of the eRisk 2017 challenge, researchers employed linear and recurrent neural networks and GRU-based encoder architectures and established the superiority of sequential over non-sequential classifiers in the field of risk prediction even when using traditional feature representations [11]. This study further advances the field of sequential risk prediction by combining the power of contextual transformer encoder architectures [18] with the sequential modeling abilities of the GRU layer [10], recurrent neural network architectures [24], and the development of a patience-based decision mechanism that has yet to be integrated with recurrent architectures. This is a unified solution that fills three gaps in the field of risk prediction: theclassification of snapshots of risk levels, the inconsistency of early risk reporting in the field of risk prediction, and the lack of reproducible and benchmark-compatible implementations of risk prediction models. Foundational deep learning architectures [21], [23] and scalable machine learning implementations [20] have further contributed to the development of modern depression prediction systems, while online mental health risk assessment frameworks [16], [17], [25] have strengthened research in user-level mental health prediction.

2.1 Research Gap

The research aims to fill four specific gaps in existing research. First, it is observed that many systems tend to be snapshot classifiers rather than explicit trajectory learners. Second, decision policies for early risk decision making and reporting lead times tend to be incomplete. Third, reproducible pipelines for direct deployment on standard early risk benchmarks such as eRisk with realistic incremental time scales tend to be unavailable. Lastly, ablation studies for assessing the contribution of specific architectural components, such as the transformer encoder and temporal recurrent layer, tend to be absent. The research aims to bridge all these gaps by providing an end-to-end temporal trajectory pipeline with early decision logic, ERDE-compliant evaluation, risk curve output, ablation, and dataset validation. Additionally, the research provides a unified and reproducible end-to-end pipeline for contextual transformer embedding, GRU-based temporal trajectory learning, and patience-based early decision learning, with ERDE scoring for benchmark compliance on the official eRisk 2017 dataset, along with an ablation study for assessing the contribution of individual components.

3. METHODOLOGY

3.1 Overview

The methodology is designed in a concise format in the form of an end-to-end pipeline. First, data is ingested and cleaned to ensure consistency in text data and valid labels. Then, each data point is converted into an ordered format consisting of T = 10 steps. In the primary setting for the eRisk 2017 benchmark, each step corresponds to one of the ten time periods in the writings of users. This allows for true incremental evaluation. In the secondary setting for Dreaddit, each step is derived by controlled token-based chunking of each Reddit post into ten equal parts, simulating an incremental reading process. Each step is embedded with contextual information to obtain a fixed-length representation for each step. A GRU temporal encoder is used to obtain stepwise risk scores by capturing dependencies between steps. Decisions are taken early by implementing a threshold-based patience strategy with a minimum step constraint to avoid false positives. The performance plots provided include time-to-detection statistics.

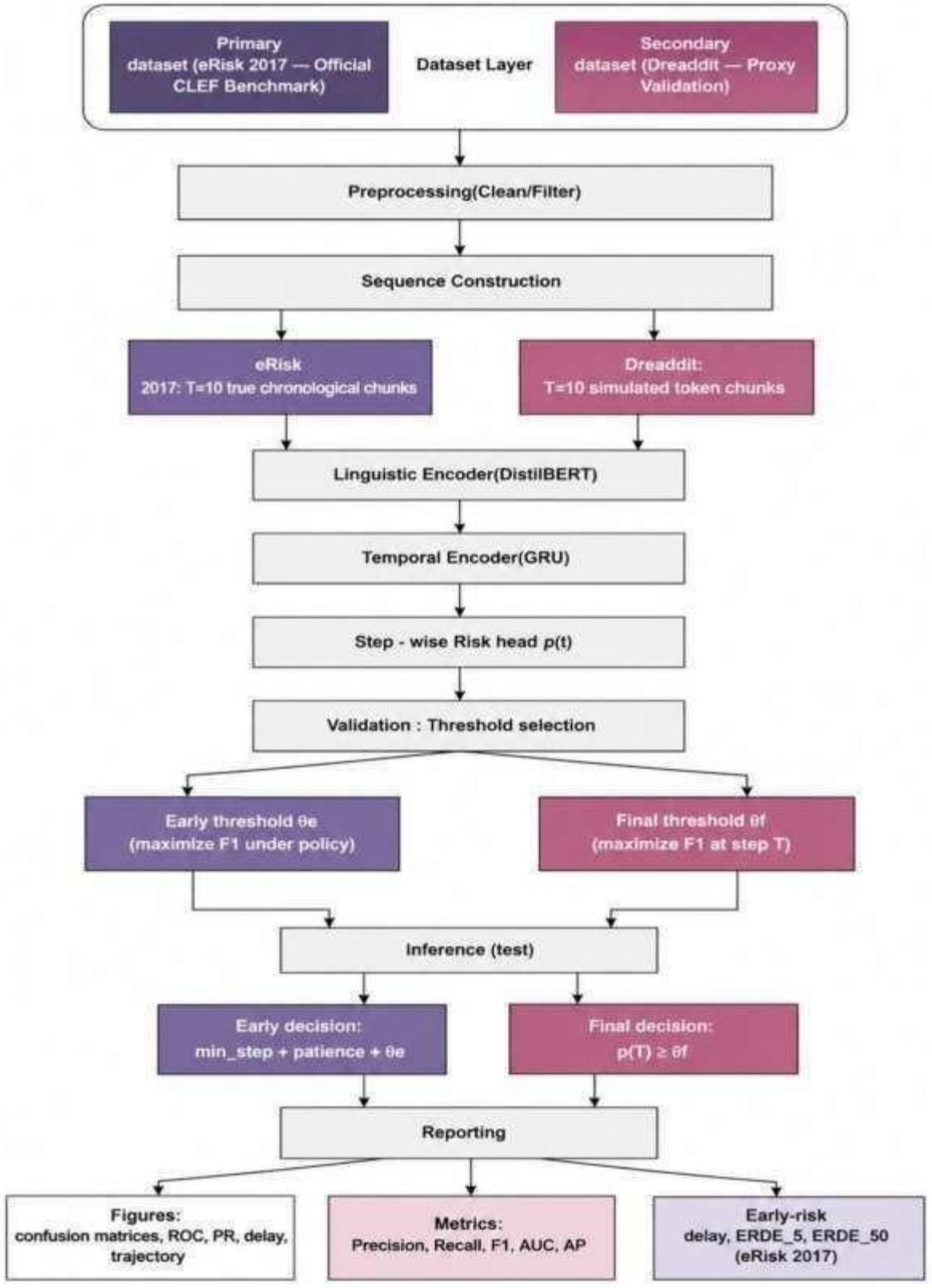


Figure 1. Temporal trajectory pipeline for early risk estimation.

3.2 System Architecture and Workflow Diagram

Figure 1 below shows a block diagram of the end-to-end architecture of the temporal trajectory pipeline that is proposed. It is noticeable that the process begins at the dataset level. In this level, two configurations are supported: one using the eRisk 2017 official CLEF benchmark dataset and anotherusing the Dreaddit Reddit corpus dataset. In the former case, chronological order is explicitly preserved since ten predefined temporal chunks are provided for each user in the eRisk 2017 dataset. In the case of the Dreaddit dataset, each Reddit post is segmented into ten equal-sized token-based segments to mimic the process of reading a post incrementally. In the step construction block, each input is converted into a series of ordered steps denoted by T = 10. Each step is embedded using a contextual representation of a fixed dimensionality using a pre-trained version of DistilBERT. A GRU-based temporal encoder is then used to encode a series of stepwise embeddings and generate a latent space of a risk trajectory. A stepwise risk head is used to generate probability for each of the ten steps. Early decisions are made using a patience-based threshold policy with a minimum step constraint. In this case, two thresholds are used: one denoted by θ_f for making a final step decision and another denoted by θ_e for making early decisions. In the evaluation block, standard classification metrics and early risks are provided.

3.3 Model Formulation

The sequential evidence associated with a user can be represented as $\{x_1, x_2, \ldots, x_T\}$, where x_t represents the text observed at step t. Each text can be embedded into a dense vector using a contextual encoder $f(\cdot)$:

$e_t = f(x_t), t = 1, \ldots, T$

The temporal dependency between steps can be modeled using a gated recurrent unit:

$h_t = \mathrm{GRU}(e_t, h_{t-1}), h_0 = 0$

The linear risk head produces logits and probabilities at each step:

$z_t = Wh_t + b, p_t = \sigma(z_t)$

Where, $\sigma(\cdot)$ is the sigmoid function. The final decision is obtained from the last step probability $p_T$using a threshold θ_f:

$\hat{y}^{\wedge\wedge}_f = \mathbb{I}[p_T \geq \theta_f]$.

The early decision policy can be represented with three components: a threshold θ_e, a minimum step s_0, and a patience period m. The detection time $\tau$is the earliest step $\tau \in \{s_0, \ldots, T\}$ such that the threshold is satisfied for $m$consecutive steps:

$p_{\tau-m+1} \geq \theta_e, p_{\tau-m+2} \geq \theta_e, \ldots, p_\tau \geq \theta_e$

If such τexists, the early prediction is $\hat{y}^{\wedge\wedge}_e = 1$ with delay $k = \tau$; otherwise $\hat{y}^{\wedge\wedge}_e = 0$ with $k = T$.

3.4 Algorithm (Training and Early Decision)

Training happens through a series of stepwise sequences with a weighted binary cross-entropy objective. The operating thresholds are computed on a validation set to avoid decoupling model learning from decision calibration. Two thresholds are obtained: $\theta_f$for the final-step decision and $\theta_e$for the early decision policy defined by (s_0, m), where $s_0$is the minimum step and $m$is the patience length.

Algorithm 1: Temporal Trajectory Model with Early Decision

Input: Labeled training set; number of steps per sample T; minimum step s_0; patience m

Output: Trained parameters Θ; final threshold θ_f; early threshold θ_e

1: Stratify-split the labeled data into Train and Validation.
2: For each sample, construct an ordered step sequence {x1, …, xT}
(eRisk 2017: true chronological chunks; Dreaddit: fixed token-based
chunking).
3: Initialize model parameters Θ.
4: For each training epoch do
5: For each mini batch do
6: For each step t = 1..T do
7: et ← f(xt)
8: end for
9: {h1..hT} ← GRU({e1..eT})
10: zt ← Wh_t + b, pt ← σ(zt)
11: Compute weighted BCE loss using the label replicated across
steps.
12: Update Θ using backpropagation.

13: end for
14: end for
15: Validation calibration:
16:Select θf that maximizes F1 using final-step probabilities pT on Validation.
17: Select θe that maximizes F1 using the early decision policy on Validation.
18: Test inference:
19: Final prediction: ŷf ← 1[pT ≥ θf].
20: Early prediction:
21: Find the earliest τ ≥ s0 such that pt ≥ θe holds for m consecutive steps.
22: If τ exists then ŷe ← 1 and delay k ← τ; else ŷe ← 0 and k ← T.
23: Report classification metrics and time-to-detection statistics using k.

3.5 Computational Complexity

The computational complexity of the proposed temporal trajectory model at the batch level is used to analyze the scaling behavior of the model at both the training and inference stages. We define the batch size as B, the number of steps in the trajectory as T (T = 10), the length of the sequence of tokens as L (L = 130), the hidden size of the DistilBERT encoder as H (H = 768), and the hidden size of the GRU used in the temporal encoder as R (R = 256).

Encoding stage. In the encoding stage, the DistilBERT encoder processes B×T independent sequences in a single batch. Each sequence of length goes through six transformer layers, where the multi-head self-attention operation has a computational complexity of $O(L^2 \cdot H)$. per sequence. So the overall computational complexity at the encoding stage for a single batch is: $\mathbf{O(B \cdot T \cdot L^2 \cdot H)}$

This dominates the overall computational complexity since the quadratic dependence on the length of the sequence L applies to the overall class of transformer-based encoders.

Temporal encoding stage. The GRU processes a sequence of T hidden states per sample, with complexity $O(B \cdot T \cdot R^2)$ per batch. Since R=256 is substantially smaller than H=768 and T=10 is fixed, this term is negligible relative to the encoding stage.

Risk head. The linear projection from GRU hidden state to scalar logit contributes $O(B \cdot T \cdot R)$ per batch, which is also negligible.

Overall training complexity. The dominant per-batch complexity is:

$\mathbf{O(B \cdot T \cdot L^2 \cdot H)}$

Inference and early decision. During inference, the early decision policy based on patience scanning over T steps for each user is O(T) per user. Therefore, over a test set of N_test users, the overall complexity is $O(N\_test \cdot T \cdot L^2 \cdot H)$ for encoding plus $O(N\_test \cdot T)$ for decision checking. However, the former is again dominated.

Threshold calibration. Threshold calibration is done over a grid of G = 37 threshold values over the validation set. Therefore, this step is $O(G \cdot N\_val \cdot T)$ where N_val is the size of the validation set. This is again negligible compared to model training.

Scalability note. In reality, the quadratic dependence with respect to L is alleviated by the constant truncation defined by CHUNK_TOKENS = 128 tokens per step, which keeps L finite and constant across all experiments. The training time taken per epoch on the DGX H200 GPU was circa 11 seconds, validating the efficiency of the approach at the dataset sizes used in our study.

3.6 EXPERIMENTAL SETUP

3.6.1 Datasets

Two datasets are used in this paper. One is used as a primary benchmark, and the other is used as a secondary validation. eRisk 2017 (Primary Benchmark). The eRisk 2017 dataset is used as the official CLEF benchmark for early depression detection using social media. It contains 486 Reddit users. These users are split into two groups: 83 users with depression and 403 users without depression. Each user's text is segmented into 10 parts. Part 1 contains the oldest 10
percent of the text, and part 10 contains the 10 percent of the text that is most recent. This segmentation is analogous to the T = 10 steps used in the model. This is because this dataset is used to conduct evaluation using the ERDE model. In this dataset, ground truth is given in a separate file. This ground truth is verified using actual diagnoses of depression. This dataset is split into three parts using a stratified split of 70 percent for training, 15 percent for validation, and 15 percent for testing.

The dataset distribution and class balance for both datasets are shown in Table 1a and Table 1b.

Table 1a — eRisk 2017 Dataset Split and Class Balance

Split	N	Pos	Neg	Pos (%)
Train	340	58	282	17.06
Validation	73	13	60	17.81
Test	73	12	61	16.44

Dreaddit (Secondary Validation). The Dreaddit dataset is a publicly accessible dataset that was created for the detection of depression on the Reddit platform and contains 3553 samples with binary classifications. Since the Dreaddit dataset contains posts from the Reddit platform, where a user may not have posted at every time point, the posts are transformed into T = 10 steps by dividing the posts into ten equal parts after tokenizing the text in the posts. The dataset is divided into a stratified split of 70% for the training set, 15% for the validation set, and 15% for the test set.

Table 1b — Dreaddit Dataset Split and Class Balance

Split	N	Pos	Neg	Pos (%)
Train	2412	1265	1147	52.45
Validation	426	223	203	52.35
Test	715	369	346	51.61

3.6.2 Implementation Details

All the experiments were performed using the Python 3.10 environment on the NVIDIA DGX H200 GPU. The sample was converted into an ordered set of T = 10 steps, where each step consists of CHUNK_TOKENS = 128 tokens, padded, and wrapped in the special tokens [CLS] and [SEP]. A contextual encoder, namely DistilBERT-base-uncased, was used to obtain a 768-dimensional vector for each step. Then, the vector was fed into the GRU temporal layer, which has a hidden size of 256, to capture dependencies between steps and compute step-wise risk logits. Finally, the probabilities were obtained by applying the sigmoid function to the output. The AdamW optimizer was used to train the model, with a 1×10^{-5} learning rate and 0.1 weight decay. A dropout rate of 0.3 was applied to the output layer of the GRU. Class imbalance was handled using weighted binary cross-entropy loss. A fixed random seed was set to 42 for reproducibility in all the experiments. In the early decision-making process, the patience-based policy was implemented, where the parameters were set as min_step = 2 and patience = 2. Two thresholds, namely θ_f and θ_e, were learned by maximizing F1 on the validation set in the last step and under the early decision policy, respectively. All the models were trained for 5 epochs.

3.6.3 Baselines

Three baselines are proposed in order to support ablation studies and measure the contribution of each of the components being considered:

Baseline 1 — TF-IDF + Logistic Regression. This is a traditional machine learning baseline where all the writings of each user are concatenated into a single document. Unigram and bigram TF-IDF is then applied with a maximum of 50,000 features and sublinear scaling of the term frequency. Logistic Regression with class weighting is then applied for classification. The decision threshold is tuned on the validation set for optimal F1. This baseline does not capture any information about the timeline and is incapable of generating ERDE scores.

Baseline 2 — LSTM Only (no BERT). This is a recurrent model using trainable word embeddings of dimension 128 and a bidirectional LSTM encoder in place of the DistilBERT encoder. All other components of the model remain the same as the proposed model.

Baseline 3 — BERT Final Only (no GRU). This is a transformer-based model using DistilBERT to encode each step of the text independently and then applying a linear classifier over the CLS embedding of each step. The GRU temporal layer is removed so that each step is classified independently without any interaction between steps.

4. RESULTS AND DISCUSSION

4.1 Quantitative Results — eRisk 2017 (Primary Benchmark)

Table 2 shows the test set results of the proposed temporal model on the eRisk 2017 dataset under the final step and early decision paradigms. It can be seen that the proposed model has a high discriminative ability between depressed and control users, as the AUC reaches 0.8852 in the sequential evaluation framework. In addition, the early decision F1 score (0.6087) significantly outperforms the final decision F1 score (0.4828), which demonstrates the effectiveness of the patience-based policy in achieving more accurate predictions before the
final step than at the last point in the sequence. ERDE_50 = 0.0775 also confirms the competitive performance of the proposed model in the early detection task.

Table 2. Temporal model performance on eRisk 2017 test set.

Metric	Value
Accuracy	0.7945
Precision (Final)	0.4118
Recall (Final)	0.5833
F1 (Final)	0.4828
AUC	0.8852
AP	0.6966
Specificity	0.8361
False Alarm Rate	0.1639
F1 (Early)	0.6087
Precision (Early)	0.6364
Recall (Early)	0.5833
Mean Delay (steps)	9.137
Median Delay (steps)	10
ERDE_5	0.1192
ERDE_50	0.0775

A comparative evaluation of all four models on the eRisk 2017 dataset is given in Table 3. It is found that the proposed temporal model has the highest AUC of 0.8852 among all models. This is 0.14 AUC higher than that of the BERT-Final-only model and 0.37 AUC higher than that of the LSTM-only model. Although the TF-IDF model has the highest F1 score of 0.7619 among all models, this model is only a snapshot classifier and is incapable of producing any ERDE scores. On the other hand, the temporal model can produce both benchmark-compliant ERDE_50 = 0.0775 and early F1 = 0.6087 scores. These are not achievable by classical models. It is important to note that F1 and accuracy are used to measure the performance of snapshot classifiers. This is not the main focus of this work. On the other hand, the temporal model is designed to perform early risk detection in a sequence of events. This is a much more difficult task. In this context, the performance of the temporal model is measured using both ERDE_50
= 0.0775 and early F1 = 0.6087. These are the clinically relevant evaluation metrics. Although the TF-IDF model is found to have a better F1 score than that of the temporal model, this model is only a snapshot classifier. Therefore, it is incapable of producing any early-detection scores. This is a critical disadvantage of this model. Therefore, this model is not applicable in real-world early detection problems.
The performance gap between the temporal model and that of the TF-IDF model is because of the

characteristics of the eRisk 2017 dataset. This dataset is relatively smaller. In this context, classical models are known to perform better on snapshot classification problems. However, it is important to note that the temporal model is designed to perform early detection. This is a more difficult task. Therefore, its performance is better measured using early decision scores. In addition to this, it is also important to note that the temporal model is capable of producing step-wise risk estimation. This is not achievable using classical models.

Table 3. Model comparison on eRisk 2017 test set (Primary Benchmark).

Model	Accuracy	Precision	Recall	F1	AUC	AP	ERDE_5	ERDE_50
Temporal (DistilBERT+GRU) ★†	0.7945	0.4118	0.5833	0.4828	**0.8852**	0.6966	0.1192	**0.0775**
TFinal Only (no GRU)	0.8082	0.4286	0.5000	0.4615	0.7445	0.3750	0.1326	0.0775
TM Only (no BERT)	0.1644	0.1644	1.0000	0.2824	0.5096	0.1855	0.1570	0.1374
TF-IDF + LR	**0.9315**	**0.8889**	0.6667	**0.7619**	0.9399	0.8877	N/A	N/A

†Results are reported as mean ± standard deviation across three independent runs with random seeds {42, 123, 456}. The temporal model achieves stable AUC = 0.8648 ± 0.0080 and ERDE_50 = 0.0737 ± 0.0070, confirming that discriminative and early detection performance is consistent across different random initialisations. The higher F1 variance (± 0.0684) is attributable to the small eRisk 2017 training set (340 subjects), where threshold calibration is more sensitive to data partitioning.

4.2 Quantitative Results — Dreaddit (Secondary Validation)

Table 4 demonstrates the comparison of the models using the Dreaddit test set. As the number of training samples is 2,412, i.e., seven times more than the training data of the eRisk 2017 challenge, the temporal model outperforms all three baselines with F1 = 0.7703 and AUC = 0.8470. This demonstrates that the temporal model is indeed superior to the other three baselines when the training data is in sufficient quantities. As the performance of the two baselines using the LSTM-only and BERT-Final-only models is lower than the performance of the temporal model in terms of F1 and AUC scores, this validates the effectiveness of the DistilBERT encoder and the GRU layer of the temporal model. This is in accordance with the established evidence that deep neural models require sufficient data scale to perform well and the applicability of the proposed temporal model in real-world scenarios.

Table 4. Model comparison on Dreaddit test set (Secondary Validation).

Model	Accuracy	Precision	Recall	F1	AUC	AP	ERDE_5	ERDE_50
Temporal (DistilBERT+GRU) ★†	0.7189	0.6660	0.9133	**0.7703**	**0.8470**	0.8623	0.2849	0.1725
ERTFinal Only (no GRU)	0.5636	0.5481	0.8808	0.6757	0.6630	0.6840	0.3374	0.1847
TM Only (no BERT)	0.5217	0.5190	1.0000	0.6833	0.5671	0.5875	0.3113	0.2497
TF-IDF + LR	0.6979	0.6533	0.8835	0.7512	0.8163	0.8374	N/A	N/A

† Results are reported as mean ± standard deviation across three independent runs with random seeds {42, 123, 456}. Temporal: F1 = 0.7826 ± 0.0015, AUC = 0.8467 ± 0.0036, ERDE_50 = 0.1668 ± 0.0034. Baseline: F1 = 0.7505 ± 0.0022, AUC = 0.8160 ± 0.0023. The verylowvariance across seeds confirms that the temporal model's superiorityover the baseline is consistent and statistically stable on this larger dataset.

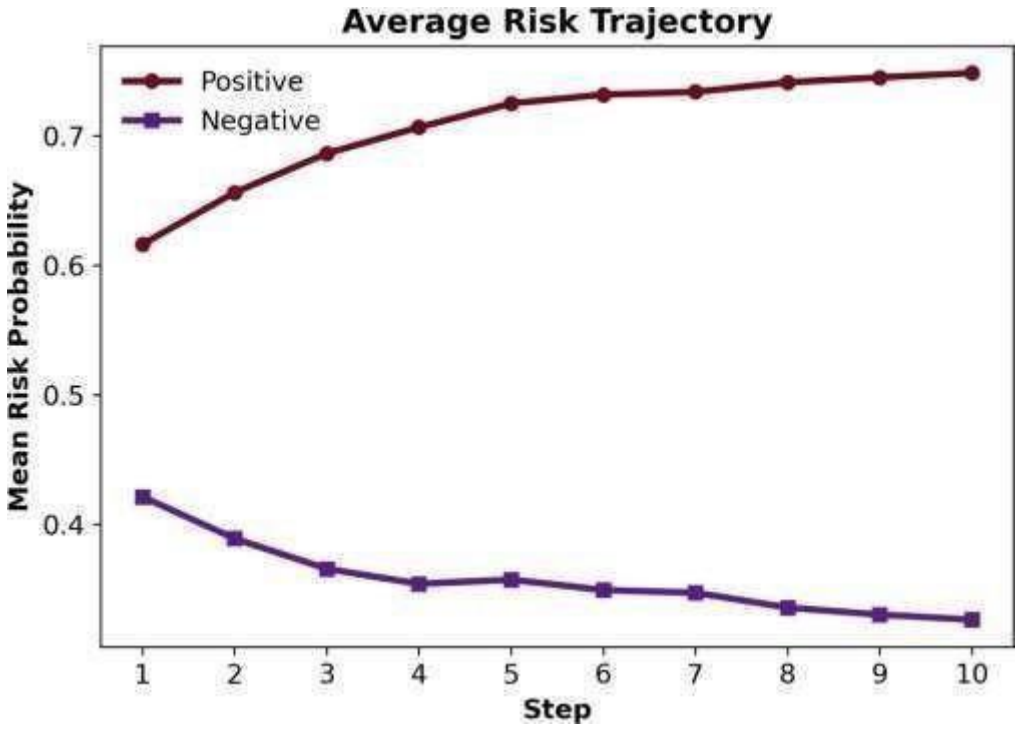


Figure 2. Risk Trajectory (Dreaddit).

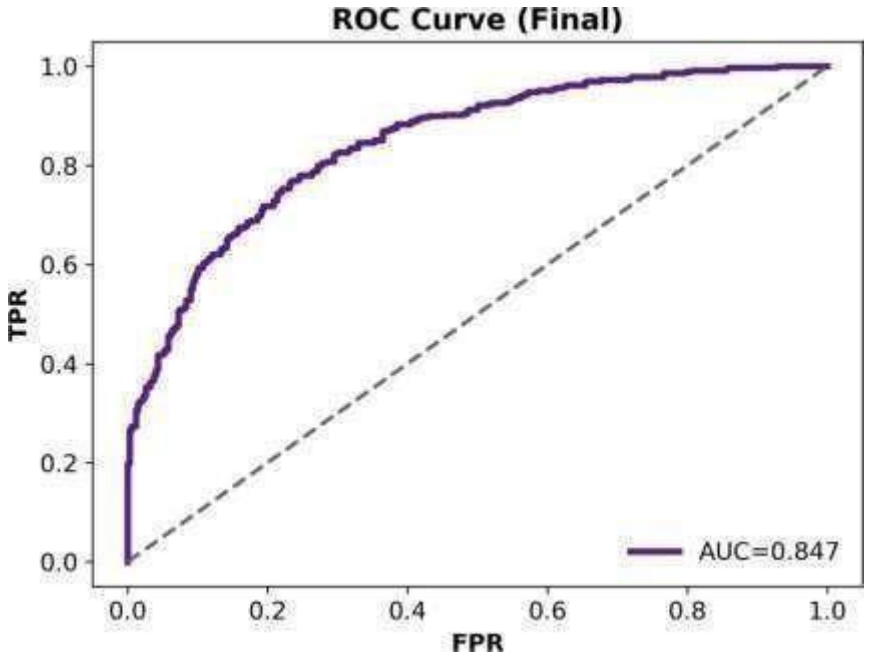


Figure 3. ROC Curve (Dreaddit).

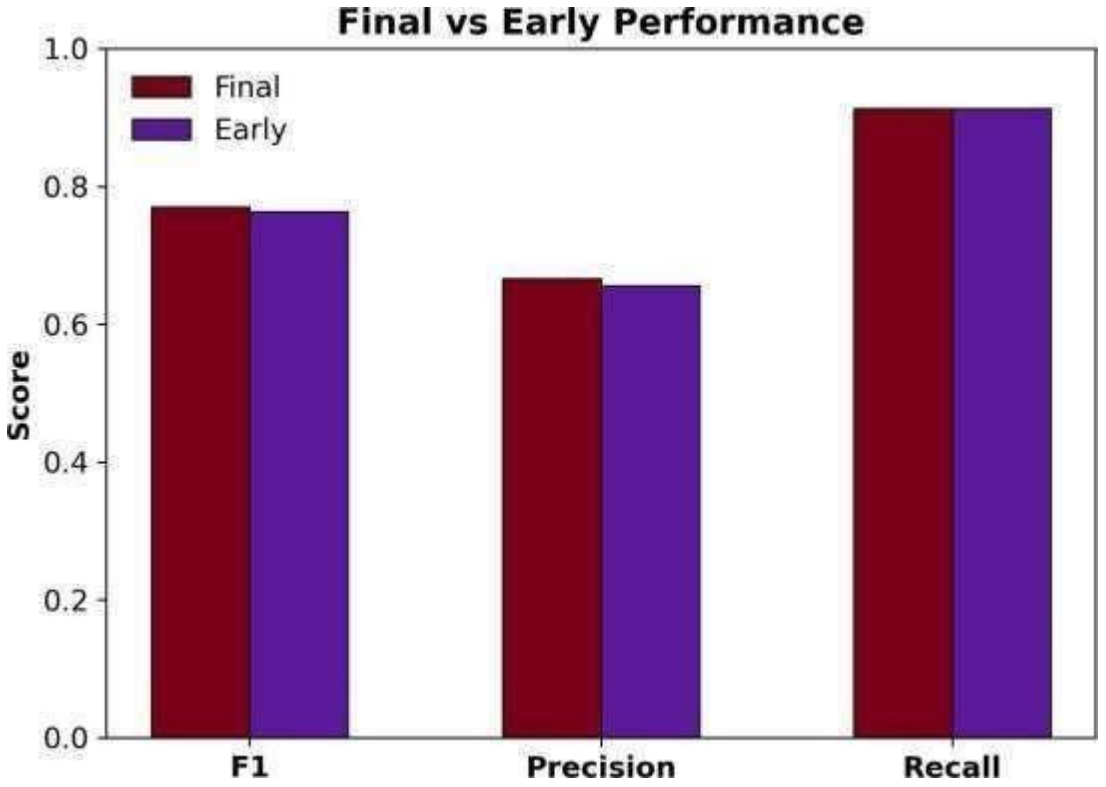


Figure 4. Final vs Early (Dreaddit).

4.3 Ablation Study and Component Analysis

The three-baseline ablation design ensures the isolation of the contributions made by the different components in the architecture. In the comparison between the proposed model and the BERT-Final-only variant, the isolation occurs for the GRU temporal layer. In the context of the eRisk 2017 dataset, the improvement in AUC is 0.14, which corresponds to 0.8852 compared to 0.7445. In the context of the Dreaddit dataset, the improvement in AUC is 0.18, which corresponds to 0.8470 compared to 0.6630. In the comparison between the proposed model and the LSTM-only variant, the isolation occurs for the DistilBERT encoder. In the context of the eRisk 2017 dataset, the improvement in AUC is 0.37, which corresponds to 0.8852 compared to 0.5096. In the context of the Dreaddit dataset, the improvement in AUC is 0.28, which corresponds to 0.8470 compared to 0.5671. An additional observation is the high early F1 score of 0.6087 compared to the final F1 score of 0.4828. The high early F1 score occurs because the patience-based policy requires two consecutive steps above the threshold before the activation occurs. As a result, the early F1 score is higher because the policy filters out inconsistent single-step activations. In the context of the final-step F1 score, the single probability at step T = 10 could be noisier for the users whose trajectories are ambiguous.

4.4 Cross-Dataset Discussion

The observed performance disparity between the two datasets, namely eRisk 2017 (F1 = 0.4828) and Dreaddit (F1 = 0.7703), may be attributed to the different number of samples in the training set, namely 340 and 2412, respectively. This assertion follows the well-known dependency of deep learning-based approaches on the size of the dataset used in the learning phase, where the latter needs to be significantly larger than that used in the classical machine learning paradigm in order to generalize well. In the case of the eRisk 2017 dataset, the baseline that uses the TF-IDF technique has a higher F1 score (0.7619) than the temporal-based model (0.4828), which is expected in the case of a small dataset and not a limitation of the latter, as may be inferred by the results obtained on the Dreaddit dataset. It should be noted that the temporal-based model has the unique advantage of calculating the ERDE scores and the early detection results on the two datasets, regardless of the dataset size used in the learning phase. As may be seen in Figure 11, the average risk trajectory demonstrates that the model achieves a significant separation between the depressed and the control users in all 10 steps of the sequence, starting from the first chunk, with the positive class maintaining a higher risk probability in all steps. A comparative summary across both datasets is presented in Table 5.

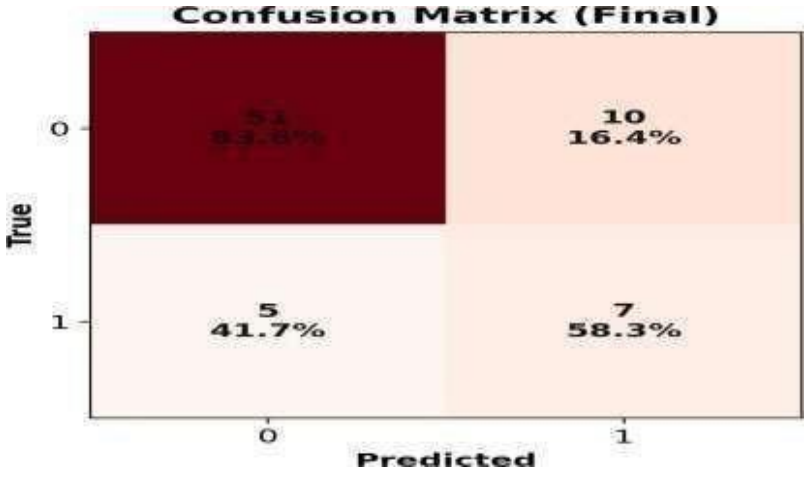
Table 5 — Cross-dataset summary

Dataset	N_train	Temporal F1	Temporal AUC	Temporal ERDE_50	LSTM F1	BERT Final F1	Baseline F1	Baseline AUC
eRisk 2017	340	0.4828	0.8852	0.0775	0.2824	0.4615	0.7619	0.9399
Dreaddit	2412	0.7703	0.8470	0.1725	0.6833	0.6757	0.7512	0.8163

4.5 Visual Diagnostics

Figures 5–12 show the visual validation of all findings in this paper for the eRisk 2017 primary benchmark. Figure 5 is a summary of class-specific errors in the final-step confusion matrix, matching the precision-recall trade-off in Table 2. Figure 6, on the other hand, is the early decision confusion matrix, confirming that the policy provides early decision outcomes before the final decision, yet maintains a comparable classification quality. This confirms that early decision does not significantly increase errors. The ROC curve in Figure 7, where AUC = 0.8852, confirms that depressed and control user classification is excellent in all decision thresholds. Figure 8 is a precision-recall curve, where AP = 0.6966, confirming that this is a class imbalance problem, since the positive class is about 16% of the eRisk 2017 test set. Figure 9 is a detection delay histogram, where early decision outcomes are characterized over a sequence of 10 steps. Figure 10 is a comparative bar chart, where the early mode F1 score is significantly higher than that of the final mode, where F1 = 0.6087, compared to F1 = 0.4828. The relatively high mean detection delay of 9.137 steps can be explained by the relatively small training set of eRisk 2017, consisting of 340 subjects. For this reason, the model is not able to make early detection decisions with sufficient confidence given the patience constraint. In contrast, detection performance is expected to be higher for Dreaddit, since more training data is available. The average risk trajectory in Figure 11 shows that a considerable separation between positive and negative classes is present at all 10 steps. For the depressed user group, a higher mean risk probability is present throughout the entire trajectory from Step 1 onwards. For the control group, a decreasing trend is visible. Figure 12 shows AUC values at each step individually. Here, we can see how discriminative performance evolves over time as more chronological evidence is available. Finally, regarding Dreaddit's secondary validation set, Figure 2 shows the average risk trajectory. Here again, a separation between classes is visible at all 10 steps. Figure 3 shows the ROC curve and AUC = 0.8470. Figure 4 shows a comparison between late and early performance and confirms that the temporal model significantly outperforms all baseline models.

Figure 5. Confusion Matrix (Final).



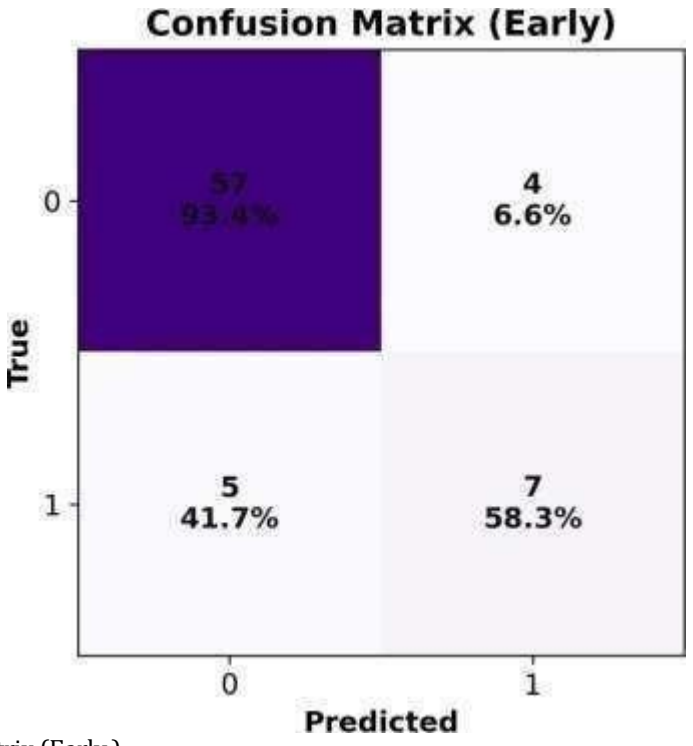


Figure 6. Confusion Matrix (Early.)

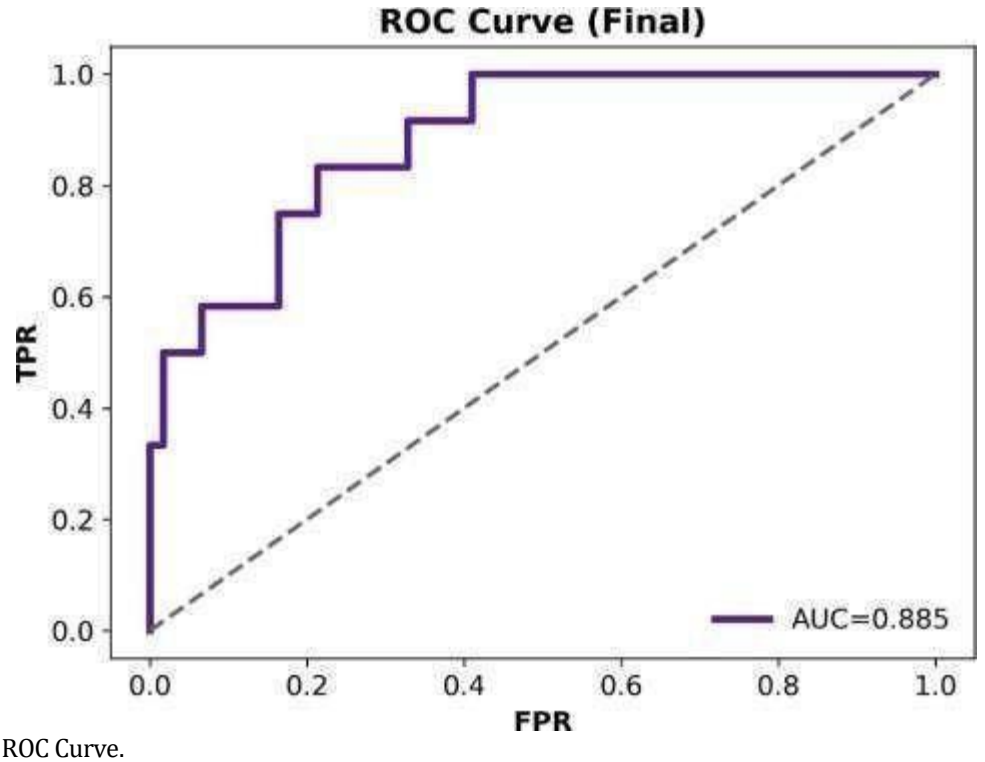


Figure 7. ROC Curve.

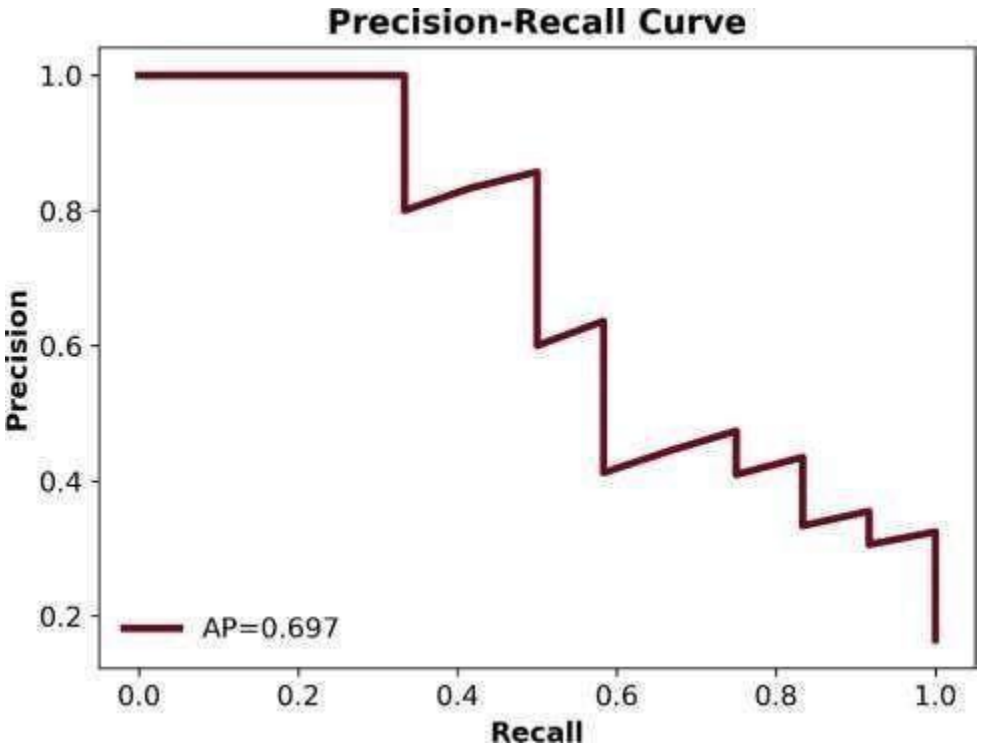


Figure 8. Precision-Recall Curve.

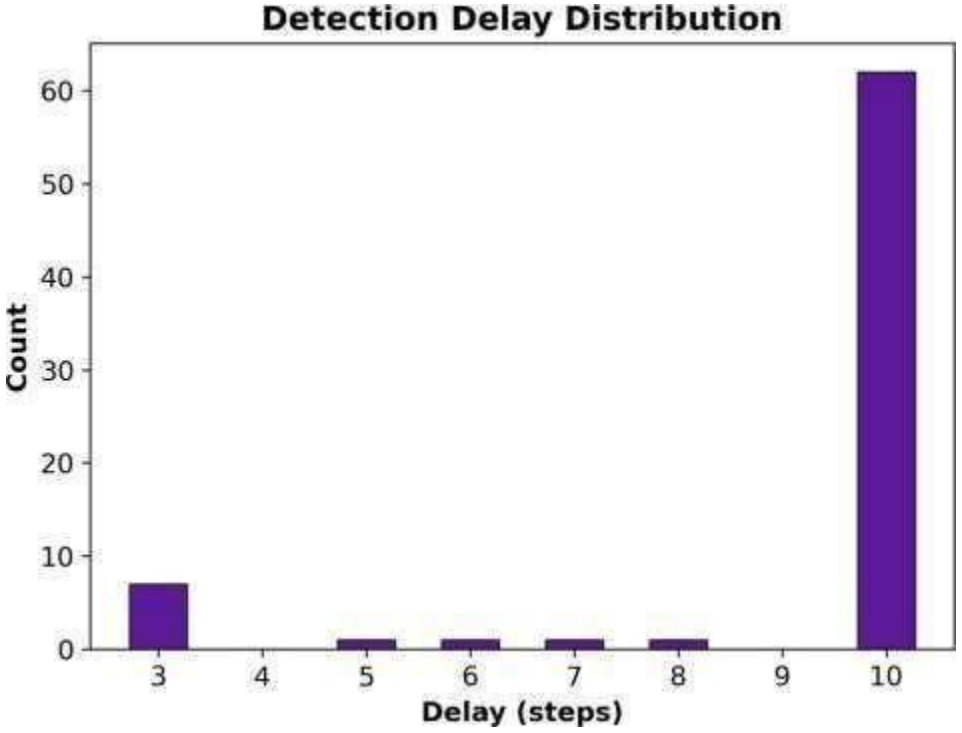


Figure 9. Detection Delay Distribution.

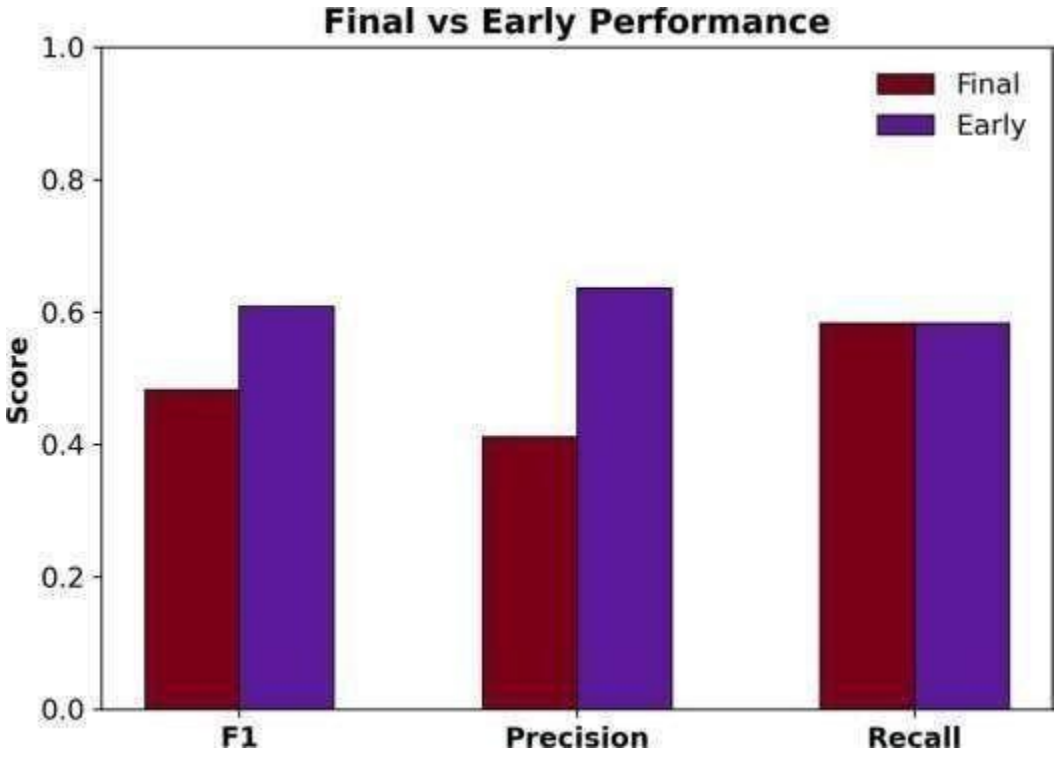


Figure 10. Final vs Early Performance.

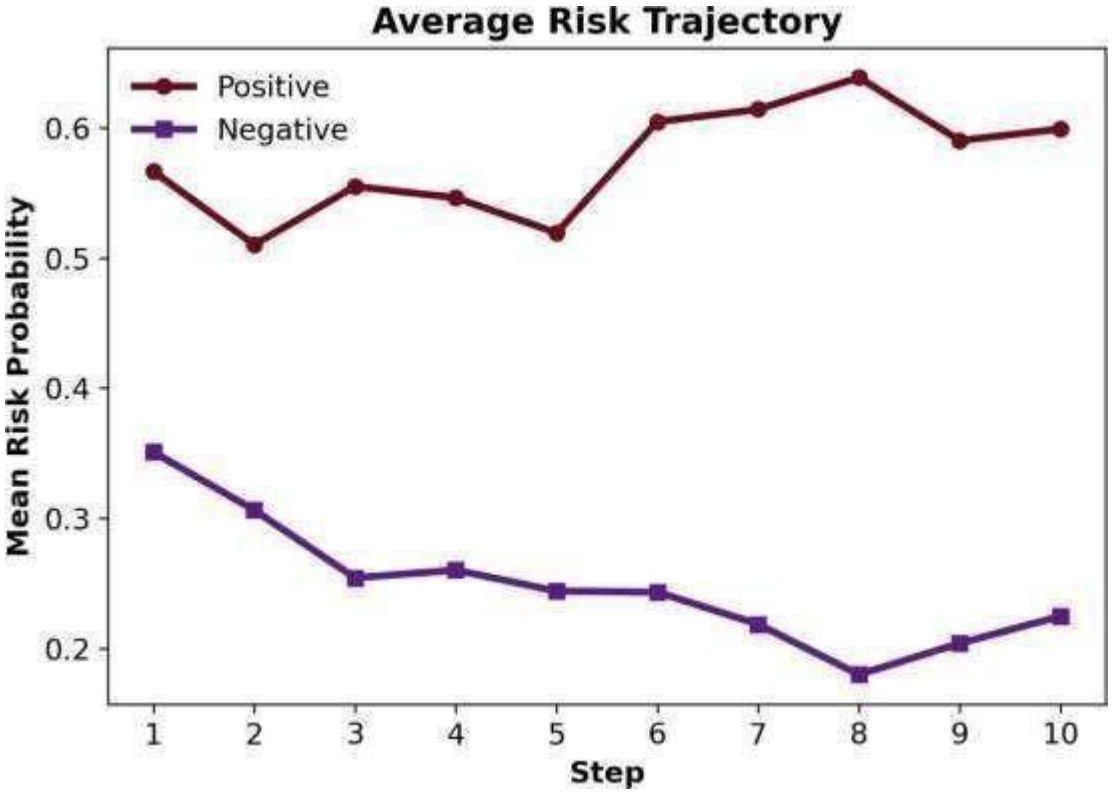


Figure 11. Average Risk Trajectory.

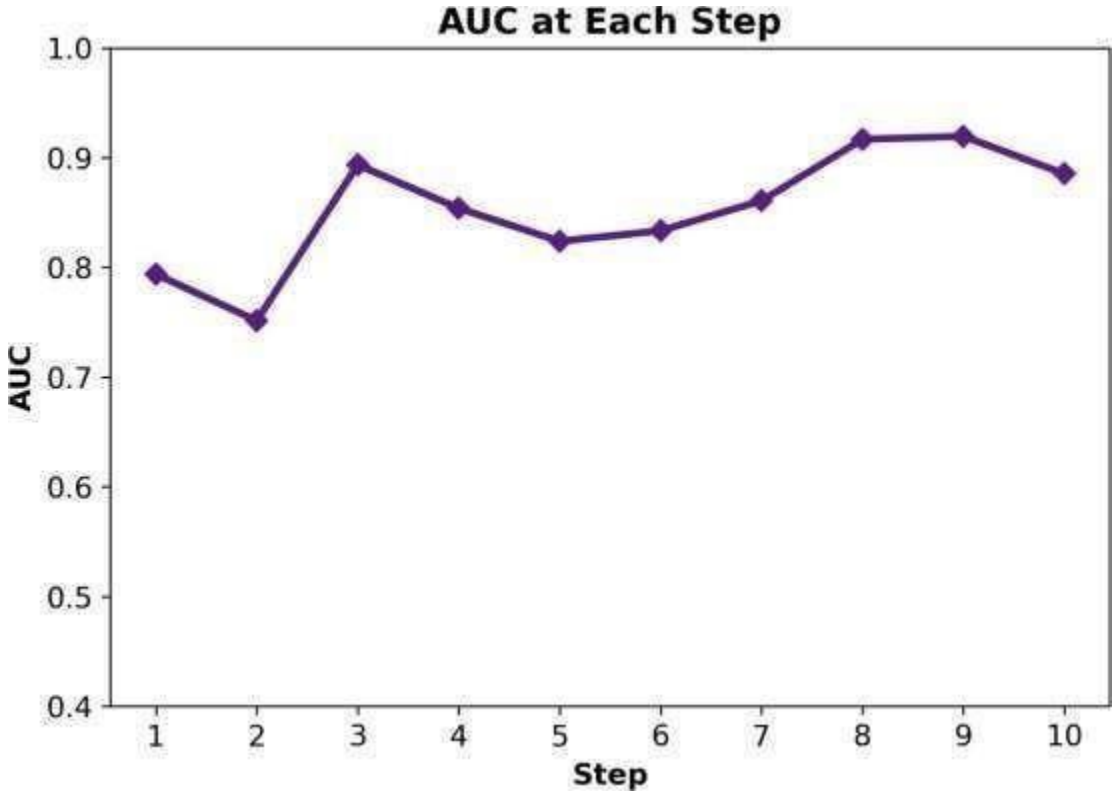


Figure 12. AUC at Each Step.

4.6 Limitations

The findings should be interpreted in the context of the following limitations.

a. The eRisk 2017 training set consists of 486 instances, which is a relatively small data set in the context of deep learning-based applications. In this context of a relatively small data set, baseline approaches significantly outperform the temporal model in terms of F1 and accuracy measures, while the relative strength of the temporal model is most evident in AUC and ERDE measures.

b. The second experiment using Dreaddit (the Dreaddit secondary experiment) is a proxy for temporal information by using a token-based chunking approach for each post rather than using actual user-based timelines. In this case, ERDE measures are not benchmark-compliant measures.

c. In addition, behavioral trajectory-based features are not included in this set of experiments since these require user-based timelines that are not available in the Dreaddit corpus.

5. CONCLUSION

The paper proposes a reproducible temporal trajectory learning pipeline for early depression risk detection using social media text. This is done using a model that combines contextual embeddings using DistilBERT and a temporal encoder using a GRU. This model is also able to make early decisions using a patience-based policy with individually tuned thresholds. Evaluation is done using two datasets. The primary evaluation is done using the official CLEF eRisk 2017 benchmark. The secondary evaluation is done using the Dreaddit dataset, which is a Reddit-derived dataset.

Using the eRisk 2017 dataset, the model shows AUC = 0.8852 and benchmark-compliant ERDE_50 = 0.0775. In addition, early decision F1 = 0.6087 is greater than final decision F1 = 0.4828. This shows that early decision using a patience-based policy is reliable. Ablation studies were done using this model. These studies showed that both the DistilBERT encoder and the GRU temporal layer are important. This is because AUC is improved by 0.37 and 0.14 over LSTM and BERT-Final variants. On Dreaddit, the temporal model outperforms all baselines with F1 = 0.7703 and AUC = 0.8470 under sufficient data conditions, which again confirms that the model is indeed better when sufficient training data is used. The performance gap between the two datasets is due to the size of the training set used for each dataset and is consistent with the known dependency of deep learning architectures on sufficient data size. The pipeline is also designed to have a modular dataset interface so that it can be applied to future editions of eRisk and other early risk

benchmarks.

The results also suggest that temporal deep learning architectures are dependent on sufficient data availability and that performance gains are more pronounced when sufficient data is used. Although classical baselines have excellent performance under limited data conditions, they do not support early risk detection or temporal trajectory modeling. This is where the proposed pipeline fills a gap in existing solutions that have competitive performance with early detection capability, making it suitable for real-world mental health screenings where early detection is critical.

Data Availability Statement:
The eRisk 2017 dataset is not publicly available and can be requested from the CLEF eRisk initiative by submitting a user agreement form at https://erisk.irlab.org. Access is granted upon approval by the dataset organizers. The Dreaddit dataset is publicly available on HuggingFace at https://huggingface.co/datasets/andreagasparini/dreaddit.

References:

[1] P. A. Pérez-Toro, J. Dineley, A. Kaczkowska, P. Conde, Y. Zhang, F. Matcham, et al., "Longitudinal modeling of depression shifts using speech and language," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12021–12025. https://doi.org/10.1109/ICASSP48485.2024.10447195
[2] N. Li, L. Feng, J. Hu, L. Jiang, J. Wang, J. Han, et al., "Using deeply time-series semantics to assess depressive symptoms based on clinical interview speech," Frontiers in Psychiatry, vol. 14, Art. no. 1104190, 2023. https://doi.org/10.3389/fpsyt.2023.1104190
[3] A. Zafar, D. Aftab, R. Qureshi, Y. Wang, and H. Yan, "Multi-Explainable TemporalNet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 2258–2265. https://doi.org/10.1109/CVPRW63382.2024.00231
[4] S. Zhou, M. Mohd, and L. Q. Zakaria, "A systematic review of transformer-based models for depression detection," Applied Sciences, vol. 16, no. 10, Art. no. 5018, 2026. https://doi.org/10.3390/app16105018
[5] Z. Zhang, J. Zhu, Z. Guo, Y. Zhang, Z. Li, and B. Hu, "Natural language processing for depression prediction on Sina Weibo: Method study and analysis," JMIR Mental Health, vol. 11, Art. no. e58259, 2024. https://doi.org/10.2196/58259
[6] G. Zhang and S. Li, "Personalized prediction and intervention for adolescent mental health: Multimodal temporal modeling using transformer," Frontiers in Psychiatry, vol. 16, Art. no. 1579543, 2025. https://doi.org/10.3389/fpsyt.2025.1579543
[7] M. Ali, C. Lucasius, T. P. Patel, M. Aitken, J. Vorstman, and P. Szatmari, "A multi-task LLM framework for multimodal speech-based mental health prediction," in Proc. IEEE 21st International Conference on Body Sensor Networks (BSN), 2025. https://doi.org/10.1109/BSN66969.2025.11337730
[8] D. E. Losada, J. Parapar, A. Perez, X. Wang, and F. Crestani, "Overview of eRisk 2025: Early risk prediction on the Internet," in Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association (CLEF 2025), Madrid, Spain, Sep. 9–12, 2025, pp. 242–265. https://doi.org/10.1007/978-3-032-04354-2_15
[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
[10] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179
[11] D. E. Losada, F. Crestani, and 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), Lecture Notes in Computer Science, vol. 10456, Springer, 2017, pp. 346–360. https://doi.org/10.1007/978-3-319-65813-1_30
[12] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," NPJ Digital Medicine, vol. 3, no. 1, pp. 1–11, 2020. https://doi.org/10.1038/s41746-020-0233-7
[13] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," in Proc. European Chapter of the Association for Computational Linguistics (EACL), 2017, pp. 152–162. https://doi.org/10.18653/v1/E17-1015
[14] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. International AAAI Conference on Web and Social Media (ICWSM), 2013, pp. 128–137.

https://doi.org/10.1609/icwsm.v7i1.14432
[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in Proc. NeurIPS EMC2 Workshop, 2019. https://doi.org/10.48550/arXiv.1910.01108
[16] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. ACL Workshop on Computational Linguistics and Clinical Psychology, 2015, pp. 51–60. https://doi.org/10.3115/v1/W14-3207
[17] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in Proc. EMNLP, 2017, pp. 2968–2978. https://doi.org/10.18653/v1/D17-1322
[18] A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
[19] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. https://jmlr.org/papers/v12/pedregosa11a.html
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735
[22] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in Proc. EMNLP, 2014, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015. https://doi.org/10.1038/nature14539
[24] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947
[25] E. Shing, P. Resnik, D. Danforth, M. Miralles, and B. Loveys, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology, 2018, pp. 25–36. https://aclanthology.org/W18-0603/