ESDM: Early Sensing Depression Model in Social Media Streams

Bichen Wang,Yuzhe Zi, Yanyan Zhao*, Pengfei Deng, Bing Qin
Harbin Institute of Technology, Heilongjiang, China
{bichenwang, yuzhezi, yyzhao, pfdeng, qinb}@ir.hit.edu.cn

Abstract

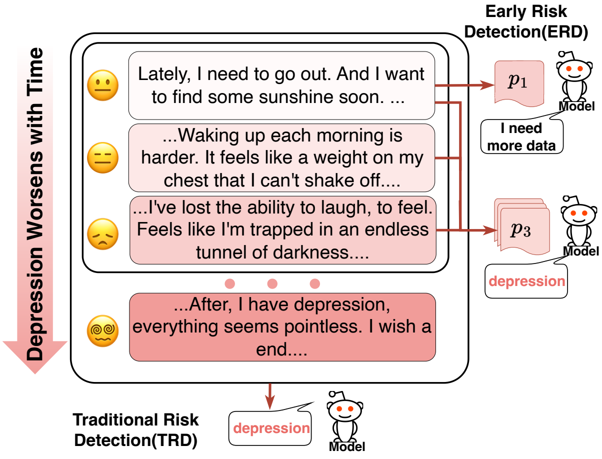
Depression impacts millions worldwide, with increasing efforts to use social media data for early detection and intervention. Traditional Risk Detection (TRD) relies on a user's complete posting history for predictions, while Early Risk Detection (ERD) seeks early detection in a user's posting history, emphasizing the importance of prediction earliness. However, ERD remains relatively underexplored due to challenges in balancing accuracy and earliness, especially with evolving partial data. To address this, we introduce the **E**arly **S**ensing **D**epression **M**odel (ESDM), which comprises two modules *Classification with Partial Information module* and *Decision for Classification Moment module*, alongside an early detection loss function. Experiments show ESDM outperforms benchmarks in both earliness and accuracy.

Keywords: depression, early detection, social media

1. Introduction

Depression affects approximately 264 million individuals globally and poses a significant public health challenge (James et al., 2018; Ferrari et al., 2013). In the U.S., nearly 15% of adults may encounter a major depressive episode during their lives (Kessler et al., 2005). To combat this, efforts worldwide focus on lessening the severe consequences of depression, with emerging strategies utilizing social media data for early detection and intervention of depression (Zhou et al., 2018; Malhotra and Jindal, 2022).

There are two kinds of tasks in depression detection: Traditional Risk Detection (TRD) and Early Risk Detection (ERD). As illustrated in Figure 1, TRD aims to make prediction based directly on complete user posting history and focuses on the accuracy of final predictions. Meanwhile, ERD treats user posting history as a data stream and aims to sequentially process the posting history to determine as early as possible whether a user is in a depression risk state while ensuring an acceptable level of accuracy. Detection delays would hinder timely interventions. If one could detect a user's depression risk at point p_3 rather than after the user develops suicidal thoughts, it provides more opportunities for timely support. Therefore, TRD using the complete posting history is more suited for retrospective analysis rather than actionable intervention (Loyola et al., 2018). In depression detection, ERD considering earliness of the prediction is of great importance since earliness implies a timely intervention. However, while ERD aligns more with practical applications, it remains a relatively underexplored domain (Zhang et al., 2022).


Early Risk Detection(ERD)
Depression Worsens with Time
p_1
Model
I need more data
p_3
Model
depression
Traditional Risk Detection(TRD)
depression
Model

Lately, I need to go out. And I want to find some sunshine soon. ...

...Waking up each morning is harder. It feels like a weight on my chest that I can't shake off....

...I've lost the ability to laugh, to feel. Feels like I'm trapped in an endless tunnel of darkness....

...After, I have depression, everything seems pointless. I wish a end....

Figure 1: The severity of untreated depression in individuals often worsens over time (Liu et al., 2021; Bernal-Morales et al., 2015). Compared to traditional approaches, ERD places emphasis on both earliness and accuracy.

The challenge of ERD involves two conflicting objectives: accuracy and earliness. Notably, there are no labels associated with individual time steps, and typically, a label characterizes an individual's complete posting history. If we have more observations, our predictions are more reliable, and vice versa. Thus, the earliness of prediction is often inversely related to accuracy, and balancing the trade-off between these two objectives makes this problem challenging. Most of previous research has utilized the complete user posting history to train classification models for TRD with golden labels, and adopted techniques such as risk windows, decision trees, and reinforcement learning for adopting ERD to decide when to classify based on model's result (Loyola et al., 2021, 2022, 2018; Zhang et al., 2022; Hartvigsen et al., 2019).

However, this transition from TRD to ERD induces inconsistencies between training and test-

* Corresponding author

LREC-COLING 2024, pages 6288–6298

ing phases. For instance, while training models may learn from data spanning years, testing often involves data from mere days or weeks. And TRD models primarily focus on the final points's prediction without considering the model's predictions with partial information at each time point. This inconsistency, compounded by differing presentations of depression on social media during its latent and apparent phases, complicates the direct application of TRD models to ERD challenges. Moreover, many models lack an end-to-end design optimized for balancing ERD's dual demands of earliness and accuracy, leading to challenges in early and precise classifications based on partial data in social media data stream.

To address these challenges, we introduce **E**arly **S**ensing **D**epression **M**odel (ESDM), which comprises two modules: the *Classification with Partial Information Module* (CPI) and the *Decision for Classification Moment Module* (DCM) to meet the two requirements for ERD, accurate classification based on partial data and early decision-making (Loyola et al., 2021, 2018; Losada et al., 2018). The CPI module leverages the accumulated post sequence up to a given point and generates initial predictions which emulates a real-world testing scenario. The DCM module, on the other hand, decides whether present information is sufficient for an immediate decision or if more posts are required. ESDM attempts to model the dynamic decision-making feature of ERD. Through considering each time step prediction and decision, we optimize the CPI's accurate partial sequence classification capability and DCM's early decision-making capacity. For effective training of these modules, we propose an end-to-end ERD learning objective, the early detection loss. It ensures the model not only focuses on prediction based on the complete user history but early and accurate prediction on partial data, serving as an end-to-end optimization objective for model's earliness and accuracy, while controlling the trade-off between them.

ESDM outperforms several baseline models on a benchmark dataset for ERD. Finally, we show how ESDM balances between earliness and accuracy; by adjusting the delay loss, we can modulate ESDM's performance between the two. [1]

- We introduce the ESDM model, specifically designed for the ERD task within social media data streams, comprising the CPI and DCM modules.
- We present the early detection loss, ensuring ESDM achieves a balance between earliness and accuracy.
- We conduct a series of experiments to showcase the advantages of ESDM. It preforms well not only in the ERD task but also in the TRD task.

2. Related Work

2.1. Depression Detection

Depression detection is challenging due to its vague nature. Depression detection often relies on content generated by individuals in clinical interviews or social media (Gratch et al., 2014; Salas-Zárate et al., 2022). Previous research has used techniques such as Linguistic Inquiry and Word Count (LIWC) and topic modeling for depression detection via social media (Pennebaker et al., 2015; Coppersmith et al., 2015). Some strategies combine visual cues from VGGNet and user engagement metrics (Simonyan and Zisserman, 2014; Zogan et al., 2021). Recently, psychological features have been integrated with neural network frameworks, especially by combining topic-related features with attention-enriched pre-trained models (An et al., 2020; Song et al., 2018). However, using psychological interview transcripts is limited due to data scarcity, so models often focus on user-centric topics correlated with depression questionnaire prompts (Rinaldi et al., 2020; Delahunty et al., 2019). Contemporary efforts also explore depression diagnosis through standardized scales, emphasizing the use of metaphors and moral language in detection (Han et al., 2022; Coll-Florit et al., 2021).

2.2. Early Risk Detection

In contrast to previous studies, ERD adopts a proactive approach by leveraging social media data streams. Over the years, the eRisk workshop has explored various contexts for identifying psychological disorders and predatory conversations (Losada et al., 2018, 2019, 2020; Parapar et al., 2022), consistently attracting academic interest. To address initial model instability leading to decision errors, a risk window-based early detection method has been proposed, where predictions are made only when the model demonstrates consistency within a specified window (Sadeque et al., 2018). With the rising volume of early detection data, a specialized incremental classifier for textual data has been introduced (Burdisso et al., 2019a,b). Researchers have also focused on enhancing existing depression models by refitting the classifier on shorter sub-sequences (Loyola et al., 2021). Additionally, scale-based solutions for identifying fixed-length features and queue-based decision-making methods have been proposed (Zhang et al., 2022).

[1] Our code will be released in https://github.com/wangyong848/Early-Depression-Detection.git

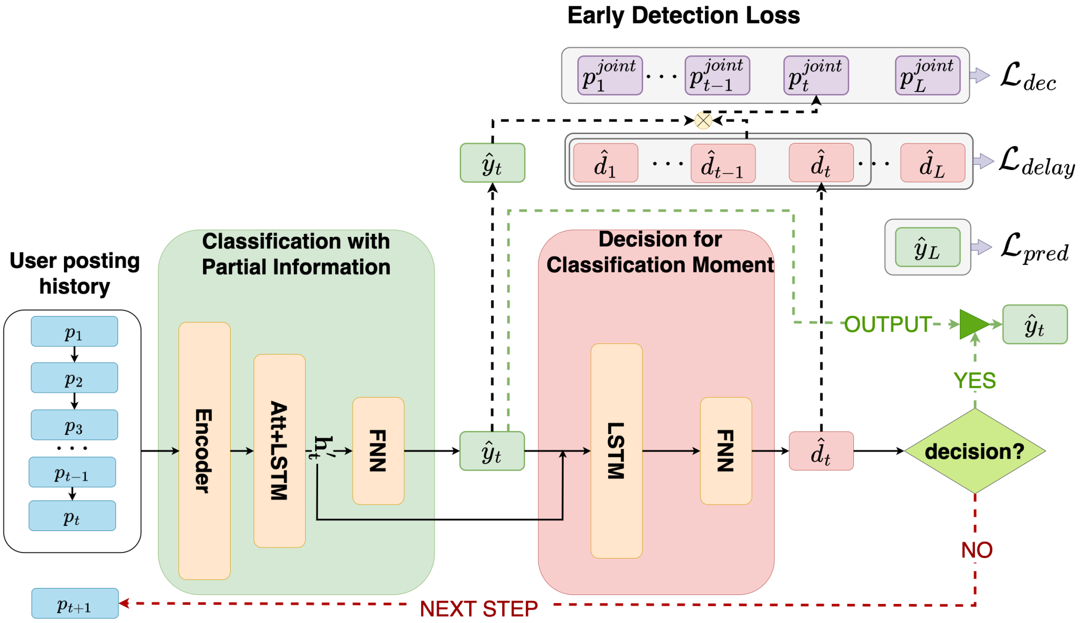


Figure 2: ESDM's overview. During inference, the red and green dashed sections determine at each step whether a decision could be made. During training, once the entire sequence is completed, we calculate our early detection loss for training. $\hat{y}_t$ is the model's current prediction, and $\hat{d}_t$ determines whether the model needs to continue. p_t^{joint} represents the joint probability that the model decides to stop at point t and makes a depression prediction.

In comparison, the majority of these solutions are primarily in the form of pipelines, while our method offers an end-to-end solution.

3. ESDM: Early Sensing Depression Model

Given a user U_i, with a posting history $P_i = \{p_{i1}, \ldots, p_{iL}\}$ where p_{ij} denotes the j-th post by U_i, our goal is to detect users with depression based on P_i. The result is represented as $\hat{y}$ which can take values 0 (indicating "not depressed ") or 1 (indicating "depressed "). Accordingly, users are classified as either depressed or not.

In the ERD, as shown in Figure 2, our ESDM runs as follows: for every step t, the partial posting history is $P_{it} = \{p_{i1}, \ldots, p_{it}\}$. At this stage, CPI generates an initial prediction $\hat{y}_t$, whereas DCM makes a decision $\hat{d}_t$. If DCM decides the prediction, $\hat{y}_t$ is treated as the final prediction. Otherwise, the model progresses to read subsequent posts. The ESDM is optimized using early detection loss functions, aiming for early and accurate detection user with depression.

3.1. Classification with Partial Information (CPI)

The primary role of the CPI module is to make an efficient predictive model that classifies the current partial sequential information up to a specific time point. The CPI derives a representation of the current partial sequence $p_{i1}, \cdots, p_{it}$ and generates a prediction denoted as $\hat{y}_t$. Moreover, in this incremental environment, unidirectionality will reduce computation and storage costs. And we integrate an attention mechanism to highlight depression-related content.

$$\mathbf{p} = f_\phi(p) \tag{1}$$

where f_ϕ represents our encoder, and we've employed BERT and $\mathbf{p} \in \mathbb{R}^{768}$. To further illuminate, we employ the LSTM to model current social media sequences, complemented with the attention mechanism to focus on content related with depression:

$$\mathbf{h}_t = LSTM_{pre}(\mathbf{h}_{t-1}, \mathbf{p}_t) \tag{2}$$

$$e_i = \mathbf{v}^T \tanh(\mathbf{W}\mathbf{h}_i + \mathbf{b}) \tag{3}$$

$$\alpha_{ti} = \frac{\exp(e_i)}{\sum_{j=1}^{t} \exp(e_j)} \tag{4}$$

$$\mathbf{h}'_t = \sum_{i=1}^{t} \alpha_{ti} \mathbf{h}_i \tag{5}$$

where $\mathbf{v}$ and $\mathbf{W}$ are learned parameters, used in computing the attention weights and $\mathbf{h}_t$, $\mathbf{h}'_t \in \mathbb{R}^h$. The CPI's prediction for the partial sequence at time step t is computed as:

$$\hat{y}_t = sigmoid(FFN(\mathbf{h}'_t)) \tag{6}$$

where $\hat{y}_t \in \mathbb{R}$ is the initial probability of depression based on the current partial sequence. Given that it is derived from a partial sequence,, the model can choose to either make decision and trust $\hat{y}_t$ or continue get posts. Therefore, the DCM module will tell us on when to make decision.

3.2. Decision for Classification Moment (DCM)

The primary role of the DCM module is determining whether the model should make decision based on existing prediction $\hat{y}_t$. The DCM module not only uses the prediction $\hat{y}_t$ but also incorporates $\mathbf{h'}_t$, it is essential to understand the representations from which predictions are made. Upon concatenation of these, the combined data is feed into the DCM module:

$$\mathbf{h}_t^{dec} = LSTM_{dec}(\mathbf{h}_{t-1}^{dec}, [\hat{y}_t : \mathbf{h'}_t]) \tag{7}$$

$$\hat{d}_t = sigmoid(FFN(\mathbf{h}_t^{dec})) \tag{8}$$

where $\hat{d}_t \in \mathbb{R}$ determines whether the model needs to make decision on this $\hat{y}_t$ or continue get next post.

3.3. Early Detection Loss for Model Training

In this section, we introduce the early detection loss for ESDM. Our objective is to ensure that, besides offering accurate predictions at final time points, the model also provides correct predictions at early time points and can make timely decisions to present these predictions as final result. In other words, through our DCM and CPI, the model can get both accurate and relatively early results. These collaborative efforts give rise to three primary training objectives that guide the training of the ESDM model:

- $\mathcal{L}_{pred}$ focuses on predictions based on the complete user history, serving as the foundation for the CPI module's ability to classify a user with depression.

- $\mathcal{L}_{dec}$ is designed to enhance the decision-making capabilities of DCM and the partial information classification abilities of CPI to ensure that ESDM can find earlier prediction and decide to output it.

- $\mathcal{L}_{delay}$ encourages DCM to make decisions as early as possible.

L_{pred} focuses on the CPI's prediction based on the user's complete posting history, for which we have a golden label. CPI needs to have a foundational ability in depression detection to be better applied to the ERD task. This loss is similar to the TRD task, we directly design it as following:

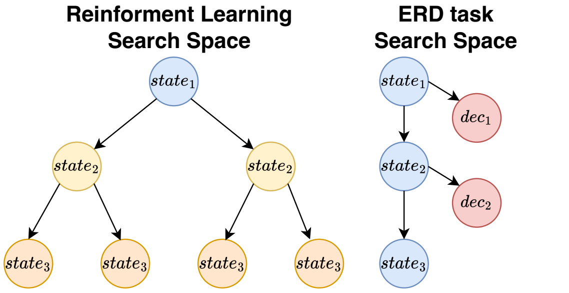


Figure 3: Differences between ERD and traditional RL environment. The right side represents the search space for ERD, while the left side represents the search space for RL environment. It can be observed that ERD may not be particularly suitable for optimization using reinforcement learning methods.

$$\mathcal{L}_{pred} = -(y \cdot \log(\hat{y}_L) + (1 - y) \cdot \log(1 - \hat{y}_L)) \tag{9}$$

where y is user's label and $\hat{y}_L$ is the CPI's prediction based on complete user posting history.

The design of $\mathcal{L}_{dec}$ is challenging because it doesn't easily fit into conventional supervised learning methods and we haven't labeled where the model should make decisions. Many researchers use reinforcement learning schemes on the similar task of unsupervised finding decision points (Yu et al., 2018; Hartvigsen et al., 2019, 2022). While some researchers employ reinforcement learning schemes in this setting, we argue that ERD might not be suited for such methods. As illustrated in Figure 3, ERD's exploration space is linear in contrast to the vast exploration space of reinforcement learning. While CPI generates predictions at each step, DCM decides whether to output this predictions. We try to compute p_t^{joint}, which means the joint probability that the model decides to stop at point t and makes a depression prediction.

$$p_t^{joint} = (\hat{d}_t \cdot \prod_{i=1}^{t-1}(1 - \hat{d}_i)) \cdot \hat{y}_t \tag{10}$$

It is diffcult to consider the predictions at each step individually. In the ERD process, the probability that the model predicts the user to be depressed is the sum of the probabilities of each point p_t^{joint}. So L_{dec} is designed as:

$$p_{all} = \sum_{t=1}^{L} p_t^{joint} \tag{11}$$

$$\mathcal{L}_{dec} = -(y \cdot \log(p_{all}) + (1 - y) \cdot \log(p_{all})) \tag{12}$$

Both the decision-making process and the prediction process are considered jointly. The predictions, $\hat{y}_t$, made at each step are taken into account, not just the final prediction, $\hat{y}_L$. Additionally, the probability of decision-making for the model is the product of the probability at that point and the joint probability of not making any decisions before that point.

To encourage models to make earlier decision points, we penalize the model's probability of not making a decision at each step. Without a cost, the result always falls at the final decision point. Hence, we introduce a simple delay loss, $\mathcal{L}_{delay}$:

$$\mathcal{L}_{delay} = \sum_{t=1}^{L} \frac{2t}{L \cdot (L+1)} (1 - \hat{d}_t) \qquad (13)$$

The factor $\frac{2}{L \cdot (L+1)}$ ensures that $\mathcal{L}_{delay}$ does not exceed 1 for each user, preventing users with more posts from receiving a larger penalty. We choose a linear cost primarily because of its inherent simplicity, ease of implementation, and fewer design intricacies. Combining the three losses gives us our final early detection loss.

$$\mathcal{L}_{ear} = \mathcal{L}_{pred} + \mathcal{L}_{dec} + \lambda \cdot \mathcal{L}_{delay} \qquad (14)$$

where λ is hyper-parameter used to balance the trade-off between earliness and accuracy in the model. We attempt to directly train the early detection model using L_{ear}. By analyzing the early detection process, we enable the model to be trained at every time step in the sequence, autonomously seeking the early decision point and classification. Concurrently, we penalize late decisions, striking a balance between earliness and accuracy.

3.4. Inference

The ESDM inference process, outlined in Algorithm 3.4, only produces a present-time depression risk prediction when the model decides to make a decision.. Once a decision is made, the model's prediction remains immutable. If the process concludes without any decision, the final prediction serves as our result.

Algorithm 1 Inference Process for ESDM

```
Require: Initial state
Ensure: Depression Prediction in Data Stream
Initialize:
for $t = 1$ to $L$ do
    Compute $\hat{d}_t$, $\hat{y}_t$
    if $\hat{d}_t$ then
        return $\hat{y}_t$
    end if
end for
return $\hat{y}_L$
```

4. Experimental Setups

4.1. Datasets

We use the eRisk-17 dataset (Losada and Crestani, 2016) in our experiment, which is adopted as the benchmark in the ERD task (Losada et al., 2017). It consists of 137 depressed users and 755 control users and is divided into training/test set with 486/406 users each. The depressed users are identified with patterns like "I was diagnosed with depression", while the control users are those active on depression subreddit but had no depression. The anchor post for identification is filtered from the dataset. This filtering strategy can prevent the direct information leakage from the self-report, which could prevent the model from learning other indirect depression signals. The statistics of the dataset are shown in the Table 1. The statistics reveals that the dataset for each user is quite extensive, encompassing about a year and a half of data.

Table 1: Train and test dataset statistics

	Train	Test
# Subjects (Dep)	83	52
# Subjects (Ctrl)	403	349
# Posts (Dep)	30,851	18,706
# Posts (Ctrl)	172,837	217,665
Avg Posts/Subj (Dep)	371.7	359.7
Avg Posts/Subj (Ctrl)	655.7	623.9
Avg Days Len (Dep)	572.7	608.3
Avg Days Len (Ctrl)	626.6	632.2

Note: Dep = Depression; Ctrl = Control; Len = Length.

4.2. Implementation Details

We initialize the hidden states of the two LSTMs with sizes 256 and 128, in that order. For the feed-forward network (FNN), a dropout rate of 0.2 is applied to both networks. To calibrate the loss function, we perform a search for the optimal λ value in the range $[10^{-1}, 5 \times 10^{-2}, \ldots, 10^{-5}]$. Ultimately, we select $\lambda = 5 \times 10^{-2}$. For training, a learning rate of 10^{-6} is chosen for the language model, while other components are set at 10^{-4}. Our model, crafted using PyTorch 1.13, is refined with the AdamW optimizer. We employ a batch size of 1 and a weight decay parameter of 10^{-3}. Training spans over 10 epochs, executed on Nvidia A100 GPUs. To mitigate natural variability, each model is run using three distinct seed values, following previous research methodologies, and the peak performance is reported. To avert numerical instability, all probability values are confined within the bounds $[10^{-7}, 1 - 10^{-7}]$.

4.3. Comparison Methods

We compare our approach with several existing methods, including those based on traditional neural networks and pretrained methods. These methods have been previously employed by researchers.

- **LR**: This approach uses TF-IDF features combined with a logistic regression classifier for prediction.
- **Feature-Enriched**: This approach integrates a suite of user-centric features, including LDA topic distributions (Blei et al., 2003), linguistic attributes from LIWC (Pennebaker et al., 2001), and metrics concerning emoji frequency.
- **BiLSTM+Attention** (Sadeque et al., 2018): Trained on a user's complete history, this approach combines the sequential data capture capability of BiLSTM with attention mechanisms.
- **Risk Window** (Sadeque et al., 2018): Trained on a user's complete history, this approach combines the sequential data capture capability of BiLSTM with attention mechanisms. A decision is made when a continuous streak of k posts produces the same result.
- **SS3** (Burdisso et al., 2019b, 2020): SS3 is an incremental classifier designed specifically for early depression detection, using an incremental learning paradigm.
- **HAN-Psych** (Zhang et al., 2022): This method integrates psychological scales into content modeling and develops an early detection mechanism using dual-layered transformers based on a queuing algorithm.
- **EARLIST** (Hartvigsen et al., 2019; Loyola et al., 2022): This technique, which incorporates reinforcement learning for early detection, has received endorsement from domain experts.

Apart from HAN-Psych, EARLIST, and the Risk Window, whose decision-making methods are already defined, we adopt their methods. For other models, we employ an immediate decision strategy: the model makes a decision when classify user as depression at anytime.

4.4. Earliness Evaluation Metrics

In addition to considering classification metrics, we introduce two early detection metrics: $erde_5$ and $erde_{50}$ (Losada et al., 2017). These two metrics are based on an exponential penalty, a variant of the sigmoid function shifted to the right. Apart from accuracy, in the scenarios that exceed the specified time limit, the model will incur heavier penalties after making a decision. The calculation formulas are as follows.

$$erde_o(k) = \begin{cases} c_{f_p}, & \text{FP} \\ c_{f_n}, & \text{FN} \\ lc_o(k) \times c_{t_p} & \text{TP} \\ 0, & \text{TN} \end{cases} \tag{15}$$

$$lc_o(k) = \frac{1}{1 + e^{(-k+o)}} \tag{16}$$

At a given decision-making point, denoted as k, the delay factor $lc_o(k) \in [0, 1]$ signifies delay-associated costs, which escalate over time. We follow the previous settings on the erisk-17 dataset, where $c_{f_p} = 0.1296$, $c_{f_n} = 1$, and $c_{tp} = 1$ (Losada et al., 2017). Consequently, $lc_o(k)$ exhibits a monotonically increasing trend with respect to k. Here, the subscript o can take on values of 5 or 50.The $erde$ metric assigns an exponential cost to the model, where if the data required by the model surpasses a certain threshold, it incurs a significant penalty. This metric underscores a model's efficiency in early decision-making.

5. Results

5.1. Comparison Result

Model	$F1_{erd}(\uparrow)$	$erde_{50}(\downarrow)$	$erde_5(\downarrow)$	$F1_{trd}(\uparrow)$
LR	0.405	0.084	0.137	0.602
Feature-Rich	0.358	0.084	0.131	0.630
BiLSTM+Att	0.562	0.096	0.124	0.629
Risk window	0.606	0.097	0.130	0.629
SS3	0.497	0.086	0.133	0.546
EARLIST	0.273	0.148	0.164	0.175
HAN-Psych	0.603	0.081	**0.107**	0.703
ESDM	**0.662**	**0.077**	0.109	**0.712**

Table 2: Main results of the experiments. The best results have been bolded. $F1_{erd}$ is the F1 score when decisions are made according to decision-making point, and $F1_{trd}$ is the F1 score of the model after considering the user's history.

As depicted in Table 2, our model consistently outperforms all baseline models. Notably, the $erde_5$ of our model might marginally trail behind HAN-Psych, yet this difference remains minimal. However, in terms of the $F1_{erd}$ and $F1_{trd}$ score, we perform better than HAN-Psych.

EARLIST's performance falls short in this task. RL, when learning ERD task from the social media text streams, sometimes doesn't performs well. EARLIST lacks substantial exploratory space, hindering the utilization of the user information. We find that not only the $F1_{erd}$ improved, but the $F1_{trd}$

also shows an increase. We believe this is because we considered the situation at every point, promoting the model's full utilization of the data.

In conclusion, the results show the performance of ESDM in the early detection of depression risks from social media data streams. Its adeptness in balancing the trade-offs between earliness and accuracy ensures timely interventions while maintaining precision.

5.2. Ablation Study

In this section, we examine the impacts of various modules in our ESDM. Given the close ties between the CPI and DCM modules to the early detection loss function, our focus down to the ablation of individual losses and their respective modules: L_{pred}, L_{dec}, and L_{delay}. The specifics of this study are as follows:

- $-L_{pred}$: The L_{pred} is removed. CPI cannot obtain the final prediction results
- $-L_{dec}$: The L_{dec} is removed. Consequently, the L_{delay} part is also removed. The DCM module is removed.
- $-L_{delay}$: The L_{delay} is removed. The model is no longer subjected to delay penalties.

  Indeed, the ablation of L_{pred} can be equated to the removal of the CPI, and likewise, L_{dec} can be perceived as the ablation of the DCM module.

Model	$F1_{erd}(\uparrow)$	$erde50(\downarrow)$	$erde5(\downarrow)$
ESDM	**0.662**	**0.077**	**0.109**
-L_{pred}	0.571	0.092	0.117
-L_{dec}	0.578	0.098	0.121
-L_{delay}	0.653	0.108	0.126

Table 3: Results of the ablation study. The ESDM results are bolded.

Table 3 shows the results of our ablation experiment. In fact, removing L_{pred} will impair the generalizability of ESDM. CPI needs to have a foundational ability in depression detection to be better applied to the ERD task, indicating that supervised learning on the complete history is still very necessary. L_{pred} ensures that the model can provide precise results after full sequence examination, preventing premature termination due to the early judgment demands of L_{dec} and L_{delay}, and instead, facilitating the identification of an appropriate stopping point. The absence of L_{dec} results in the model's inability to pinpoint reliable decision points, causing it to regress to a LSTM-Att archetype. On the other hand, the removal of L_{delay} seems to compromise the model's earliness in decision-making. These losses and their corresponding modules represent the capabilities required for an early detection model.

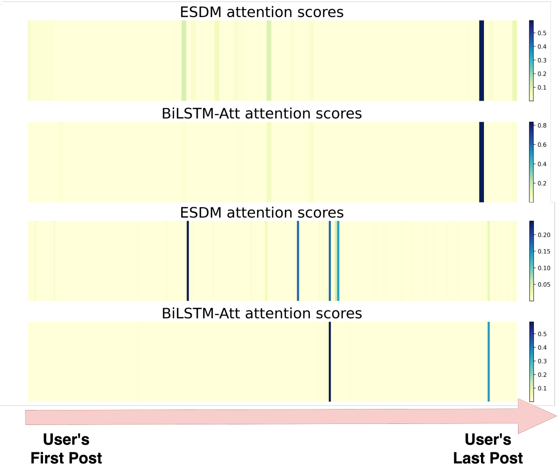


Figure 4: Attention scores for ESDM and BiLSTM-Att. When presented with complete user history, there's a noticeable difference in the attention distribution between ESDM and BiLSTM-Att, as illustrated by depression sample examples.

6. Analysis for ESDM

The ESDM exhibits better performance in ERD tasks. In this section, we aim to demonstrate some reasons contributing to this, examining both the CPI and DCM perspectives. We will demonstrate how our approach encourages the CPI to focus on some earlier posts and show how the DCM module get a balance between earliness and accuracy.

6.1. CPI: Prioritization of Earlier Posts

To demonstrate ESDM's ability to focus on earlier posts within a user's history, we compare its attention visualization with that of a BiLSTM-Att model trained directly on the entire user history. When the entire user history is shown, we observe that the CPI's attention scores tend to point towards earlier posts, allowing the CPI to rely more heavily on posts that appear earlier in the sequence. This comparison is specifically chosen due to the structural similarities between BiLSTM-Att and ESDM, which highlights the unique contribution of our work. As illustrated in Figure 4, the heatmap suggests that ESDM has a more extensive attention span across the user's history, particularly focusing on some earlier posts. This observation aligns with our expectations.

We also compute the model's average attention

Model	Num	Decision Posts
ESDM	37	...when something like this happens, it's completely natural to want to make amends. ...
HAN-BERT	103	...As to why I don't want to exist, I couldn't tell you. ...
BiLSTM-Att	100	...i need help. i just feel like i don't want to exist. ...
ESDM	26	...Headaches throughout most of the day ... I'd feel kinda OK after waking up ...
HAN-BERT	84	...usually not to bad unless it's a particularly bad depression day for me ...
BiLSTM-Att	95	...I have frequent depression, and some experience with other mental health stuff ...

Table 4: Sample decisions posts from various models.

position to understand its focal point.

$$Avg\ Attention\ Position = \sum_{t=1}^{L} t \cdot \alpha_t \quad (17)$$

For ESDM, the average attention position is 52.7, while for BiLSTM-Att, it is 141.2. This demonstrates that ESDM can focus features earlier. In the following section, we will analyze the decision-making timing of the DCM module.

6.2. DCM: Temporal Decision Dynamics

In this section, we discuss the differences between ESDM's decision-making and various baselines, as well as the impact of the balance with earliness and accuracy.

As shown in the Table 4, the decision points chosen by the other two models indicate situations where users are already in communication with therapists, or they have clearly stated their depression, even to the point of contemplating suicide. By the time these risks of depression are detected, they have already advanced to a severe stage. However, ESDM can detect a user with depression at a much earlier stage. ESDM captures signs in the early stages of depression when the user has already shown some unstable tendencies but has not yet escalated to a severe level. This provides an opportune moment for timely intervention.

As shown in the Figure 5, one can observe the model's trade-off between earliness and accuracy. As the temporal penalty intensifies, the model tends to make earlier predictions. Interestingly, the F1 value does not always decrease, nor does it reach its optimal result. On the contrary, the F1 value shows a slight decline. We believe that to some extent, the temporal penalty prompts the model to focus on the early features of depression. An appropriate temporal penalty can actually benefit the model in producing correct results. However, if the temporal penalty is too severe, the model becomes overly aggressive, making it difficult to make accurate judgments. The model achieves a certain balance around 10^{-2}.

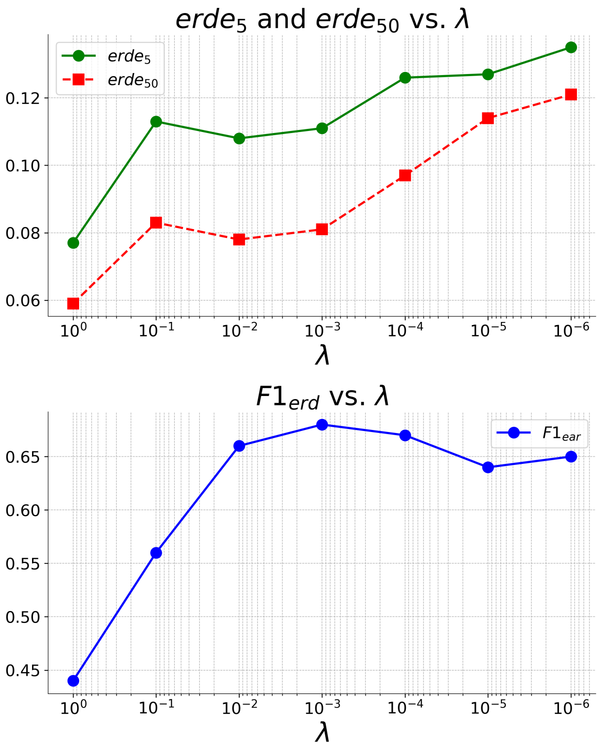


Figure 5: As the time penalty coefficient changes, the variation of F1 and erde in ESDM reflects how the model balances between earliness and accuracy.

7. Conclusion

The field has made good progress in detecting depression risks traditionally (TRD), but there's not much work done in catching them early (ERD), which is more helpful. We made a new model called the Early Sensing Depression Model (ESDM) to do just that. ESDM is like a smart tool that knows when to raise an alarm about someone's mental health, even with little information, and it doesn't rush or wait too long. Our tests showed that ESDM works better than other tools. As we keep using the internet more, tools like ESDM help use online info for good, making sure people get help when they need it.

8. Broader Impact and Ethical Considerations

The application of models like ESDM in early depression detection could revolutionize mental health interventions and may potentially save countless lives. Detecting depression risks at an earlier stage could assist medical professionals, educators, and concerned family members in providing timely support. Furthermore, such models can be instrumental in public health campaigns and policies aimed at reducing the global burden of depression.

However, with such potential benefits, there are also broader impacts and ethical considerations:

8.1. Ethical Considerations

Our research on depression risk may raise certain ethical concerns. The data used in our study are acquired from publicly shared datasets shared by other researchers. In order to protect individuals' privacy, all social media data underwent strict anonymization procedures by the data providers before being used. We comply with relevant ethical guidelines and legal regulations, ensuring that there is no risk of privacy violations during the research process. The classification is not intended as a diagnostic tool, but rather a risk estimate for individual users that can then be used to support monitoring and support for users.

8.2. Positive Outcomes

- **Timely Intervention**: One of the core strengths of the ESDM model is its focus on Early Risk Detection (ERD). By identifying depressive tendencies at their nascent stages, timely and appropriate interventions can be applied, potentially preventing a further decline in mental health or even life-threatening situations.

- **Support to Mental Health Professionals**: ESDM serves as a valuable tool for psychiatrists, therapists, and counselors. It can guide them in diagnosing and treating patients more effectively by providing data-driven insights into a patient's mental state.

- **Awareness and Education**: The incorporation of such technology can lead to broader public awareness about the importance of early detection in mental health. As more individuals recognize the capabilities of tools like ESDM, there's potential for increased self-reflection and proactive measures towards seeking help.

8.3. Negative Outcomes and Mitigation Strategies

- **Accuracy and False Positives**: While ESDM is designed for accuracy, no model is infallible. False positives may label someone as at-risk when they are not, which could lead to unwarranted interventions and psychological distress

- **False Negatives**: Conversely, false negatives are equally concerning, potentially overlooking individuals who genuinely need intervention and support.

- **Over-reliance**: The broader public, educators, and even healthcare professionals should not solely rely on such models but use them as one tool among many in comprehensive mental health assessments.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key RD Program of China via grant 2021YFF0901602 and the National Natural Science Foundation of China (NSFC) via grants 62176078 and 62236004.

Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1078–1089, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Blandina Bernal-Morales, Juan Francisco Rodríguez-Landa, and Frank Pulido-Criollo. 2015. Impact of anxiety and depression symptoms on scholar performance in high school and university students. *A fresh look at anxiety disorders*, 225.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019a. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019b. Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2020. τ-ss3: a text classifier with dynamic n-grams for early risk detection over text streams. *Pattern Recognition Letters*, 138:130–137.

Marta Coll-Florit, Salvador Climent, Marco Sanfilippo, and Eulàlia Hernández-Encuentra. 2021. Metaphors of depression. studying first person accounts of life with depression published in blogs. *Metaphor and Symbol*, 36(1):1–19.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Fionn Delahunty, Robert Johansson, and Mihael Arcan. 2019. Passive diagnosis incorporating the PHQ-4 for depression and anxiety. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 40–46, Florence, Italy. Association for Computational Linguistics.

Alize J Ferrari, Fiona J Charlson, Rosana E Norman, Scott B Patten, Greg Freedman, Christopher JL Murray, Theo Vos, and Harvey A Whiteford. 2013. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS medicine*, 10(11):e1001547.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Thomas Hartvigsen, Walter Gerych, Jidapa Thadajarassiri, Xiangnan Kong, and Elke Rundensteiner. 2022. Stop&hop: Early classification of irregular time series. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 696–705.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. 2019. Adaptive-halting policy network for early classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 101–110.

Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858.

Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602.

Jin Liu, Yiming Fan, Ling-Li Zeng, Bangshan Liu, Yumeng Ju, Mi Wang, Qiangli Dong, Xiaowen Lu, Jinrong Sun, Liang Zhang, et al. 2021. The neuroprogressive nature of major depressive disorder: evidence from an intrinsic connectome analysis. *Translational Psychiatry*, 11(1):102.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 343–361, Cham. Springer International Publishing.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR*

Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, pages 340–357. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*.

Juan Martín Loyola, Sergio Burdisso, Horacio Thompson, Leticia C Cagnina, and Marcelo Errecalde. 2021. Unsl at erisk 2021: A comparison of three early alert policies for early risk detection. In *CLEF (Working Notes)*, pages 992–1021.

Juan Martín Loyola, Marcelo Luis Errecalde, Hugo Jair Escalante, and Manuel Montes y Gomez. 2018. Learning when to classify for early text classification. In *Computer Science–CACIC 2017: 23rd Argentine Congress, La Plata, Argentina, October 9-13, 2017, Revised Selected Papers 23*, pages 24–34. Springer.

Juan Martín Loyola, Horacio Thompson, Sergio Burdisso, and Marcelo Errecalde. 2022. Unsl at erisk 2022: Decision policies with history for early classification.

Anshu Malhotra and Rajni Jindal. 2022. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, page 109713.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2022. Overview of erisk 2022: early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 233–256. Springer.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, University of Texas, UT Faculty, Austin, Texas.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18, Online. Association for Computational Linguistics.

Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Rafael Salas-Zárate, Giner Alor-Hernández, María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Maritza Bustos-López, and José Luis Sánchez-Cervantes. 2022. Detecting depression signs on social media: a systematic literature review. In *Healthcare*, volume 10, page 291. MDPI.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Keyi Yu, Yang Liu, Alexander G Schwing, and Jian Peng. 2018. Fast and accurate text classification: Skimming, rereading and early stopping.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2022. Psychiatric scale guided risky post screening for early detection of depression. *arXiv preprint arXiv:2205.09497*.

Lina Zhou, Dongsong Zhang, Christopher C Yang, and Yu Wang. 2018. Harnessing social media for health information management. *Electronic commerce research and applications*, 27:139–151.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. *DepressionNet: Learning Multi-Modalities with User Post Summarization for Depression Detection on Social Media*, page 133–142. Association for Computing Machinery, New York, NY, USA.