TUA1 at eRisk 2022: Exploring Affective Memories for Early Detection of Depression

Xin Kang[1], Rongyu Dou[1] and Haitao Yu[2]

[1]Tokushima University, 2-1, Minamijyousanjima, Tokushima, 770-8506, Japan
[2]University of Tsukuba, 1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan

Abstract

This paper describes the participation of the Tokushima University A1 (TUA1) group in the Early Detection of Depression task at the CLEF eRisk 2022 Lab. We propose a Time-Aware Affective Memories (TAM) network for early detection of the user depression risk, based on the stream of user postings on the Internet. The TAM network regularly maintains an *enriched* memory of the affective state for each user with a Time-Aware LSTM (T-LSTM) model. The embedding of the affective memory and that of the new post are then integrated through a Transformer Decoder for predicting the user's depression risk. To encourage early detections of the depression risk, we propose a latency penalty to the risk predictions during training. The model raises a risk *decision* based on the binary classification result and estimates a risk-ranking *score* based on the difference between the positive and negative probabilities. Our experimental results show that the proposed affective memory is effective in Early Detection of Depression and achieve two state-of-the-art results in the ranking-based Early Detection of Depression evaluation.

Keywords

Time-aware affective memory, affective state, latency penalty, Early Detection of Depression

1. Introduction

Early risk prediction on the Internet (eRisk) has been a long-running Lab at CLEF [1, 2, 3, 4, 5, 6], which aims at exploring the early detection technologies to predict potential risks in the Internet users' health and safety. In this year, the Early Detection of Depression task at the CLEF eRisk 2022 Lab [6] focuses on predicting the depression risk in users based on their social media postings. A user is depression-positive if an explicit mention of being diagnosed with depression was made by the user [1, 2]. By observing the posts of a user from the very beginning, a detection system needs to raise the risk *decision* as early as possible if the user is depression-positive and to estimate a risk-ranking *score* indicating the level of depression.

The early studies of language usage in depression patients [7, 8, 9, 10, 11] suggest that depression and language usage are internally correlated, while the recent psychological studies of depression [12, 13] indicate that depression is indeed a complex emotional state and highly associates with several negative emotions [14, 15], such as sad and anxiety. These findings

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
kang-xin@is.tokushima-u.ac.jp (X. Kang); c502047004@tokushima-u.ac.jp (R. Dou); yuhaitao@slis.tsukuba.ac.jp (H. Yu)
0000-0001-6024-3598 (X. Kang)

have inspired recent studies to explore linguistic features [16, 17, 18, 19, 20, 21], emotions, and sentiments [22, 23, 24] in user posts for detecting depression and several related mental disorders, such as suicide ideation [25].

We extend these studies by exploring the history of user affective states, based on the connection between depression and the long-term negative affects reflected in one's posts, that is, the difficulty of removing negative feelings from one's working memory [15, 26]. We consider affective state as the embedding of user emotion in a post, which is retrieved by a pre-trained DistilBERT-Emotion model. A Time-Aware Affective Memories (TAM) network is proposed to maintain the memory of an Internet user's affective state, which gets update with the user's latest affective state and the time interval $\Delta\tau_t$ between the user's latest (τ_t) and last (τ_{t-1}) postings. This affective memory is fed together with the semantic information of the latest post to a Transformer Decoder, and TAM uses the decoded information to predict a user's depression risk.

To encourage early detection of the depression risk, we propose a latency penalty that penalizes the latency of the *first-positive* predictions for the depression-positive users. Our initial experiment suggests that latency penalty is effective for reducing the Early Risk Detection Error (ERDE) score for the Early Detection of Depression task.

The rest of this paper is arranged as follows. Section 2 briefly reviews the recent depression detection studies. Section 3 depicts the TAM network and the latency penalty for the Early Detection of Depression task. Section 4 details our submissions to the task and presents the results. Section 5 concludes our work.

2. Related Work

Early Detection of Depression task was firstly proposed by Losada and Crestani [27], in which researchers built a test collection on depression and language and proposed ERDE for systematically evaluating early detection algorithms in accuracy and latency of depression-positive predictions. The task received 30 and 45 contributions in CLEF 2017 [1] and CLEF 2018 [2], respectively.

In Early Detection of Depression at CLEF 2017, Trotzek et al. [28] explored various linguistic and meta features, such as personal and possessive pronouns, past tense verbs, word *I*, and text readability measures. These features were combined with the n-gram feature for training a logistic regression classifier. This work won the best precision, F1, and ERDE_5 scores by averaging the logistic probabilities obtained from the above classifier and from a paragraph-embedding based logistic regression classifier. In Early Detection of Depression at CLEF 2018, Trotzek et al. [29] achieved the best ERDE_{50} and F1 scores, by setting a threshold to the logistic regression probabilities. Funez et al. [30] employed a Flexible Temporal Variation of Terms (FTVT) approach, which utilized a sequential information about the variation of terms among different post chunks. This approach obtained the best ERDE_5 in the same task. Paul et al. [31] employed an Ada Boost classifier with the Bag of Word (BoW) features and got the best precision score.

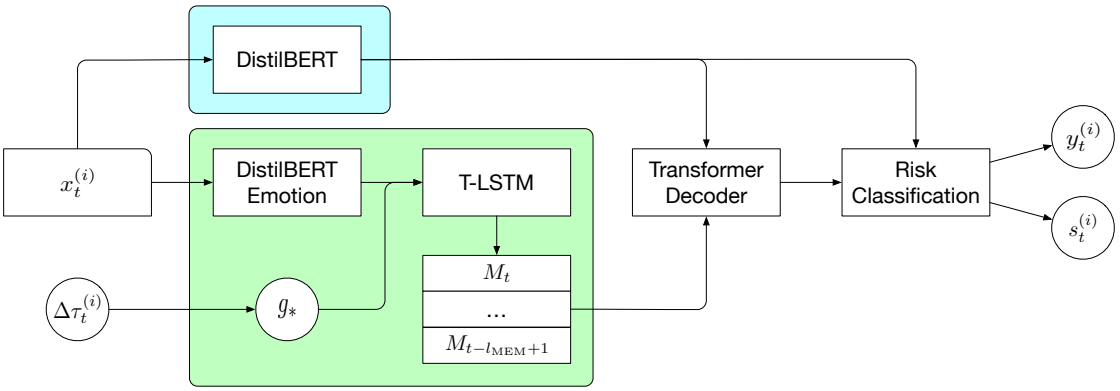

Figure 1: Overview of the Time-Aware Affective Memories (TAM) network for Early Detection of Depression. Modules for affective processing and semantic processing are indicated in the light green and light blue squares, respectively. An affective state is stored in the T-LSTM network for each user.

3. TAM Network for Early Detection of Depression

3.1. Time-Aware Affective Memories Network

To explore the history of user affective states for Early Detection of Depression, we propose a Time-Aware Affective Memories (TAM) network as shown in Fig. 1. TAM is composed of an affective processing module and a semantic processing module, which are indicated in the light green and the light blue squares, respectively.

First, the **affective processing module** expects the latest post $x_t^{(i)}$ from user i at step t and the time interval $\Delta\tau_t^{(i)}$ between the user's latest and last postings as input. We concatenate the title and body of a post into $x_t^{(i)}$, with the user-sensitive and task-insensitive information replaced with special tokens[1]. The time interval $\Delta\tau_t^{(i)}$ is given by

$$\Delta\tau_t^{(i)} = \tau_t^{(i)} - \tau_{t-1}^{(i)}, \tag{1}$$

where $\tau_t^{(i)}$ and $\tau_{t-1}^{(i)}$ are the time logs of the user's latest and last postings.

Second, the user's emotion in post $x_t^{(i)}$ is mapped into an affective state $A_t^{(i)}$ based on a pre-trained DistilBERT Emotion classification model DistilBERT_E [2]. The mapping is given by

$$A_t^{(i)} = \varphi\left(\text{DistilBERT}_E(x_t^{(i)})\right), \tag{2}$$

where the affective state $A_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ corresponds to a φ-pooled activation of the pre-classification layer in DistilBERT_E with input $x_t^{(i)}$, $\varphi(\cdot)$ corresponds to either a mean-pooling or a CLS-pooling among the first dimension of a tensor, and d_{BERT} is the DistilBERT model dimension. DistilBERT_E is pre-trained on an English Twitter Emotion dataset [32], which

[1] User-sensitive Email addresses and phone numbers are replaced with ⟨EMAIL⟩ and ⟨PHONE⟩, and task-insensitive numbers and currency symbols are replaced with ⟨NUMBER⟩ and ⟨CUR⟩, respectively with clean-text.

[2] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

classifies user postings into *joy*, *love*, *surprise* , *sadness*, *anger*, and *fear*. The pre-trained DistilBERT is slightly inferior to that of BERT in emotion classification but is over two times faster in processing speed.

Third, the affective states $A^{(i)}$ of user i is remembered by a Time-Aware LSTM (T-LSTM) network [33]. T-LSTM takes the affective state $A_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ for the current post $x_t^{(i)}$ as the first input and discounts its internal affective memory in $C \in \mathbb{R}^{d_{\text{MEM}}}$ with the time interval $\Delta\tau_t^{(i)} \in \mathbb{R}_{>0}$ as the second input. In the following description we omit the user index i for abbreviation. Given the internal memory $C_{t-1} \in \mathbb{R}^{d_{\text{MEM}}}$ and the hidden state $h_{t-1} \in \mathbb{R}^{d_{\text{MEM}}}$ at the last step $t-1$ as well as inputs $A_t^{(i)}$ and $\Delta\tau_t^{(i)}$ at the latest step t, T-LSTM updates its internal memory and hidden state by

$$
\begin{aligned}
C_{t-1}^S &= \tanh(W_d C_{t-1} + b_d) && \text{(Short-term memory)} \\
\hat{C}_{t-1}^S &= C_{t-1}^S g_*(\Delta\tau_t) && \text{(Discounted short-term memory)} \\
C_{t-1}^L &= C_{t-1} - C_{t-1}^S && \text{(Long-term memory)} \\
C_{t-1}^A &= C_{t-1}^L + \hat{C}_{t-1}^S && \text{(Adjusted previous memory)} \\
f_t &= \sigma(W_f A_t + U_f h_{t-1} + b_f) && \text{(Forget gate)} \\
i_t &= \sigma(W_i A_t + U_i h_{t-1} + b_i) && \text{(Input gate)} \\
o_t &= \sigma(W_o A_t + U_o h_{t-1} + b_o) && \text{(Output gate)} \\
\tilde{C}_t &= \tanh(W_c A_t + U_c h_{t-1} + b_c) && \text{(Candidate current memory)} \\
C_t &= f_t C_{t-1}^* + i_t \tilde{C}_t && \text{(Current memory)} \\
h_t &= o_t \tanh(C_t), && \text{(Current hidden state)}
\end{aligned}
$$

where $W_d \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{MEM}}}$ and $b_d \in \mathbb{R}^{d_{\text{MEM}}}$ are parameters for decomposing the memory. $W_f, W_i, W_o, W_c \in \mathbb{R}^{d_{\text{BERT}} \times d_{\text{MEM}}}$, $U_* \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{MEM}}}$, and $b_f, b_i, b_o, b_c \in \mathbb{R}^{d_{\text{MEM}}}$ are parameters for calculating the forget, input, output gates and the candidate current memory, respectively. g_* is a set of discount functions that monotonically decrease with the time interval $\Delta\tau_t$. We employ two discount functions g_{slog} and g_{flex} for the detection of depression task, in which g_{slog} is reciprocal to the logarithm of interval seconds

$$
g_{\text{slog}}(\Delta\tau) = 1/\log(\Delta\tau + \epsilon), \tag{3}
$$

with a hyper-parameter ϵ of 1.0, and the g_{flex} is a flexible power function of the interval seconds inspired by [34]

$$
g_{\text{flex}}(\Delta\tau) = \frac{q_1}{a\Delta\tau} + \frac{q_2}{1 + (\Delta\tau/b)^c}, \tag{4}
$$

with trainable parameters $q_1, q_2, a, b, c \in \mathbb{R}$.

Last, a linear layer is employed to map the T-LSTM hidden state $h_t^{(i)} \in \mathbb{R}^{\mathbb{R}^{d_{\text{MEM}}}}$ to an affective memory $M_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ by

$$
M_t^{(i)} = W_M h_t^{(i)} + b_M, \tag{5}
$$

with parameters $W_M \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{BERT}}}$ and $b_M \in \mathbb{R}^{d_{\text{BERT}}}$. To enrich the memorization of a user's affective states for TAM, we concatenate the most recent l_{MEM} affective memories by

$$
\hat{M}_t^{(i)} = \text{Concat}(M_{t-l_{\text{MEM}}+1}^{(i)}, \ldots, M_t^{(i)}), \tag{6}
$$

where $\hat{M}_t^{(i)} \in \mathbb{R}^{l_{\text{MEM}} \times d_{\text{BERT}}}$ is the *enriched* affective memory.

The **semantic processing module** takes the latest post writing $x_t^{(i)}$ from user i at time step t as input, which is similar as the affective processing module, and encodes it to a semantic embedding $S_t^{(i)}$ with a pre-trained DistilBERT model[3] by

$$S_t^{(i)} = \text{DistilBERT}(x_t^{(i)}), \tag{7}$$

where $S_t^{(i)} \in \mathbb{R}^{l_{\text{TEXT}} \times d_{\text{BERT}}}$ corresponds to the activation of the pre-classification layer in DistilBERT, l_{TEXT} corresponds to the length of $x_t^{(i)}$, and d_{BERT} is the DistilBERT model dimension.

To integrate the *enriched* affective memory $\hat{M}_t^{(i)}$ and the semantic embedding $S_t^{(i)}$ in TAM, we employ a Transformer Decoder network as shown in 1. We denote $\hat{M}_t^{(i)}$ and $S_t^{(i)}$ as A and B for illustrating the integration mechanism as below. A Transformer Decoder is a multi-head cross-attention architecture, each head of which makes queries for elements from an input sequence A and retrieves new values from a reference input sequence B, based on the element-wise similarity between A and B. Specifically, the cross-attention $\text{MultiHead}(A, B)$ is concatenated by $\text{head}_1, \ldots \text{head}_H$ with

$$\text{MultiHead}(A, B) = \text{Concat}(\text{head}_1, \ldots \text{head}_H)W^O, \tag{8}$$

$$\text{head}_h = \text{Attention}(AW_h^Q, BW_h^K, BW_h^V), \tag{9}$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\intercal}{\sqrt{d_2}})V, \tag{10}$$

where $A \in \mathbb{R}^{n \times d_1}$, $B \in \mathbb{R}^{m \times d_1}$ are sequences of n and m embeddings and d_1 is the embedding dimension. To empower the attention mechanism, A and B are first mapped from the d_1-dimensional space to query $Q_h \in \mathbb{R}^{n \times d_2}$, key $K_h \in \mathbb{R}^{m \times d_2}$, and value $V_h \in \mathbb{R}^{m \times d_2}$ in a larger d_2-dimensional space through linear projection with parameters $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d_1 \times d_2}$, and h is the index of attention heads. Each $\text{head}_h \in \mathbb{R}^{m \times d_2}$ is then calculated by the Attention function with the corresponding query, key, and value as the input. Last, the concatenated attention head is mapped from $m \times d_2$ back to d_1 dimension with a projection parameter $W^O \in \mathbb{R}^{Hd_2 \times d_1}$.

We propose to integrate the affective memory and semantic embedding with either one Transformer Decoder by

$$H_t^{(i)} = \text{mean}(\text{MultiHead}(\hat{M}_t^{(i)}, S_t^{(i)})), \tag{11}$$

or two Transformer Decoders by

$$\begin{aligned} H_t^{(i)} = &\text{mean}\left(\text{MultiHead}\left(\text{mean}(\hat{M}_t^{(i)}), S_t^{(i)}\right)\right) + \\ &\text{mean}\left(\text{MultiHead}\left(\varphi(S_t^{(i)}), \hat{M}_t^{(i)}\right)\right), \end{aligned} \tag{12}$$

[3] https://huggingface.co/distilbert-base-uncased

where $\mathrm{mean}(\cdot)$ indicates a mean-pooling in the first dimension of a tensor while $\varphi(\cdot)$ corresponds to either a mean-pooling or a CLS-pooling. Both decoding strategies render an integration $H_t^{(i)} \in \mathbb{R}^{d_{\mathrm{BERT}}}$.

The depression probability $p_t^{(i)}$ and its logit $\gamma_t^{(i)}$ are predicted by a Risk Classification network, based on $H_t^{(i)}$ and a φ-pooled semantic embedding $\varphi(S_t^{(i)})$. Specifically, the concatenation of $H_t^{(i)}$ and $\varphi(S_t^{(i)})$ is passed through a linear layer with layer normalization and ReLU activation, a dropout layer, and a final classification layer of the Risk Classification network. The outputs are a *score* $\hat{s}_t^{(i)}$ that indicates the level of depression

$$\begin{aligned} \hat{s}_t^{(i)} &= p_t^{(i)} - (1 - p_t^{(i)}) \\ &= 2p_t^{(i)} - 1, \end{aligned} \tag{13}$$

and a risk *decision* $\hat{y}_t^{(i)}$

$$\hat{y}_t^{(i)} = 1\{\gamma_t^{(i)} > 0\}, \tag{14}$$

where $1\{\cdot\}$ is an indicator function.

Besides the stepwise risk classification, we employ a score accumulation technique [35] that accumulates the historical risk scores for the current *score* by

$$\tilde{s}_t^{(i)} = \sum_{t'=1}^{t} \hat{s}_{t'}^{(i)}, \tag{15}$$

and predict the risk *decision* by

$$\tilde{y}_t^{(i)} = 1\left\{\tilde{s}_t^{(i)} > \mathrm{median}\left(\tilde{s}_{[1:t]}^{(i)}\right) + \gamma \mathrm{MAD}\left(\tilde{s}_{[1:t]}^{(i)}\right)\right\}, \tag{16}$$

where $\tilde{s}_{[1:t]}^{(i)}$ is a list of the accumulated scores for user i up to time step t and $\mathrm{median}(\cdot)$ renders the median value of a list. The MAD function is given by

$$\mathrm{MAD}(\tilde{s}_t^{(i)}) = \mathrm{median}\left(\left|\tilde{s}_{[1:t]}^{(i)} - \mathrm{median}\left(\tilde{s}_{[1:t]}^{(i)}\right)\right|\right), \tag{17}$$

which evaluates the Median Absolute Deviation of the accumulated scores $\tilde{s}_{[1:t]}^{(i)}$.

3.2. Latency Penalty

We propose a latency penalty ψ that penalizes TAM for the latency of the *first-positive* predictions, in terms of the depression-positive users. The latency penalty for user i at time step t is given by

$$\begin{aligned} &\psi\left(y^{(i)}, \gamma_t^{(i)}, \gamma_{\max(t)}^{(i)}, t; \alpha, o\right) = \\ &\qquad \sigma(\gamma_t^{(i)}) \cdot y^{(i)} \cdot lc\left(t \cdot \tanh\left(\alpha \cdot \mathrm{ReLU}\left(\gamma_t^{(i)}\right) \cdot \mathrm{ReLU}\left(-\gamma_{\max(t)}^{(i)}\right)\right); o\right), \end{aligned} \tag{18}$$

where $y^{(i)} \in \{0, 1\}$ is the ground truth label, $\gamma_t^{(i)} \in \mathbb{R}$ is the current predicted logit, $\gamma_{\max(t)}^{(i)} = \max_{t'=1}^{t-1} \gamma_{t'}^{(i)}$ is the maximum logit up to $t-1$, and $t \in \mathbb{Z}$ indicates the current time step. α and o are two hyper-parameters, respectively, which control the latency sensitivity and the time step at which the latency cost grows quickly as described below. σ is the sigmoid function. The latency cost function lc is first proposed in the ERDE metric [27], which is given by

$$lc(t; o) = 1 - \frac{1}{1 + \exp^{t-o}}, \tag{19}$$

with input t denoting the *latency* step of a true-positive prediction. The latency cost $lc \in (0, 1)$ monotonically grows with the *latency* step t and grows the most quickly at the step o with a latency cost of 0.5. In practice, o is usually set to 5 and 50, the latter of which is employed for training the proposed TAM network.

In Eq. 18, we obtain the *latency* of the *first-positive* prediction for user i through a series of neural activation functions of the sequence of logit predictions $\gamma_{[1:t]}^{(i)}$. Specifically, $\mathrm{ReLU}(\gamma_t^{(i)})$ renders a positive value $\gamma_t^{(i)}$ if the logit with respect to the latest (t) posting is positive, and renders 0 otherwise. Similarly, $\mathrm{ReLU}(-\gamma_{\max(t)}^{(i)})$ renders a positive value $-\gamma_{\max(t)}^{(i)}$ if all logits up to the last ($t-1$) posting are negative, and renders 0 otherwise. We scale their product with the latency sensitivity $\alpha = 10{,}000$ and feed the result to $\tanh(\cdot)$. The output turns to be an indicator that takes a value close to 1 if the model renders *a positive prediction* for the latest posting for user i and *all-negative predictions* before that, while takes the value of 0 otherwise. By multiplying the latest time step t with the indicator, we obtain the step of *first-positive* prediction, that is the *latency*, and feed it to the latency cost function in Eq. 19. The latency penalty ψ is finally given by the product of the depression probability $\sigma(\gamma_t^{(i)})$, the ground-truth label $y^{(i)}$, and the latency cost lc.

We add the latency penalty in Eq. 18 to a cross-entropy loss to produce the final training target for Early Detection of Depression by

$$\begin{aligned}\ell(y, \gamma; \alpha, o) = &\sum_{t=1}^{T} \sum_{i=1}^{N} \\ &- \left(y^{(i)} \log \sigma(\gamma_t^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(\gamma_t^{(i)}))\right) + \psi\left(y^{(i)}, \gamma_t^{(i)}, \gamma_{\max(t)}^{(i)}, t; \alpha, o\right),\end{aligned} \tag{20}$$

where N and T are the number of users and the number of time steps in the training data, respectively.

4. Experiment

The training data of Early Detection of Depression at the CLEF eRisk 2022 Lab [6] consists of the training and test data of CLEF eRisk 2017 Lab and the test data of CLEF eRisk 2018 Lab. The details can be found in Table 1.

The test data of Early Detection of Depression at the CLEF eRisk 2022 Lab [6] consists of 1,400 users. The posts of these users are accessible in an interactive manner during the test

Table 1
Number of positive and negative users in the training data of Early Detection of Depression at the CLEF 2022 Lab.

	Positive	Negative
2017	135	752
2018	79	741
Total	214	1,493

Table 2
Distinctive configurations of the submitted models.

Configuration	TUA1#0	TUA1#1	TUA1#2	TUA1#3	TUA1#4
Balance Strategy	Balance	All	Balance	All	Balance
$\varphi_{\mathrm{DistilBERT}}$	Mean	CLS	Mean	CLS	N/A
$\varphi_{\mathrm{DistilBERT-Emo}}$	CLS	Mean	CLS	Mean	N/A
Max Memory Len	30	1	30	1	Full
Discount Function	g_{slog}	g_{flex}	g_{slog}	g_{flex}	N/A
Decoder Num	1	2	1	2	N/A
Score Accumulation	False	False	True	True	True

phase, that is, the server only replies one post per-user at step t after receiving the depression predictions for all users at step $t-1$. Posts at step 0 from all users are accessible at the very beginning.

We submit five groups of risk *decisions* and risk *scores* for 2,000 steps in this interactive manner, which takes around 16.5 hours. Among all participants in Early Detection of Depression, our system turns to be *the most efficient*.

The distinctive configurations of the submitted models are shown in Table 2. Specifically, Balance Strategy indicates the way of selecting positive and negative users from the training data, for which *Balance* indicates that as many as the positive users are randomly selected from the negative set while *All* indicates that all users are utilized. $\varphi_{\mathrm{DistilBERT}}$ and $\varphi_{\mathrm{DistilBERT-Emo}}$ indicate a *Mean*-pooling or a *CLS*-pooling for $\varphi(\cdot)$ in Eq. 12 and Eq. 2, respectively. Max Memory Len corresponds to l_{MEM}, which is the length of *enriched* affective memory $\hat{M}$. Discount Function indicates the utilization of either g_{slog} or g_{flex} for discounting the short-term memory $\hat{C}^S$. Decoder Num specifies the number of Transformer Decoders in the TAM network for integrating the affective memory $\hat{M}$ and the semantic embedding S. Score Accumulation indicates predicting the depression scores and risk decisions by either accumulating the historical risk scores or not. TUA1#0 to TUA1#3 corresponds to the TAM-based models with distinctive configurations, while TUA1#4 is a SS3-based model [35]. Configurations which are not applicable to the model are denoted as *N/A*. To avoid making reckless risk *decisions*, we halt the positive predictions by producing all-zero decisions in the first two time steps for all models.

Table 3 shows the decision-based evaluation results. First, we find that Score Accumulation in the TAM-based models obtains similar decision-based evaluation scores, which is possibly because that the TAM network has already maintained a long-term memory of the affective

Table 3
Decision-based evaluation for the Early Detection of Depression task. Results obtained by our models and the best performing models on each metric are included.

Model	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
TUA1#0	0.155	0.806	0.260	0.055	0.037	3.0	0.922	0.258
TUA1#1	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1#2	0.155	0.806	0.260	0.055	0.037	3.0	0.992	0.258
TUA1#3	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1#4	0.159	0.959	0.272	0.052	0.036	3.0	0.992	0.270
CYUT#2	0.106	0.867	0.189	0.056	0.047	**1.0**	**1.000**	0.189
LauSAn#0	0.137	0.827	0.235	0.041	0.038	**1.0**	**1.000**	0.235
LauSAn#4	0.201	0.724	0.315	**0.039**	0.033	**1.0**	**1.000**	0.315
BLUE#2	0.106	**1.000**	0.192	0.074	0.048	4.0	0.988	0.190
NLPGroup-IISERB#0	0.682	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**
Sunday-Rocker2#0	0.091	**1.000**	0.167	0.080	0.053	4.0	0.988	0.165
Sunday-Rocker2#4	0.108	**1.000**	0.195	0.082	0.047	6.0	0.981	0.191
SCIR2#3	0.316	0.847	0.460	0.079	**0.026**	44.0	0.834	0.383
E8-IJS#0	**0.684**	0.133	0.222	0.061	0.061	**1.0**	**1.000**	0.144

states through T-LSTM as well as an *enriched* affective memory. Next, TUA1#0 and TUA1#2 achieve better Precision, F1, $ERDE_{50}$ and $F_{latency}$ scores than TUA1#1 and TUA1#3, which indicates a long affective memory and a balanced training data could be helpful for improving the decision predictions in TAM. Our results also suggest the importance of exploring language usage patterns for predicting the depression decisions. Last, it is reasonalbe to speculate that halting positive predictions for the first two time steps could be an important factor that reduces the latency-sensitive metric scores, such as $ERDE_5$, $ERDE_{50}$, $latency_{TP}$, and $F_{latency}$, in our result.

Table 4 shows the ranking-based evaluation results. First, the ranking-based decisions of TUA1#0 and TUA1#2 render the state-of-the-art results in P@10 and NDCG@10 based on only 1 user post. The result suggests that the TAM network with a long affective memory could effectively recognize the users' depression risk at a very early state. It also implies that taking the decision-halting strategy off from TAM might render better decision-based evaluation results. Next, TUA1#1 obtains better results than TUA1#3, which indicates that Score Accumulation might not be necessary for the ranking-based prediction in TAM. TUA1#0 and TUA1#2 generally obtain better P@10, NDCT@10, NDCG@100 scores for 1 post, 100 posts, 500 posts, and 1,000 posts, which suggests that long affective memory and balanced data are also helpful in improving the ranking-based predictions for TAM. Last, the TAM-based models significantly outperform the SS3-based model in terms of the ranking-based metrics.

5. Conclusion

In this paper, we propose a Time-Aware Affective Memories (TAM) network with a latency-penalized cross-entropy loss for Early Detection of Depression at the CLEF eRisk 2022 Lab. Both decision- and ranking-based evaluation results indicate that affective state is an important

Table 4
Ranking-based evaluation for the Early Detection of Depression task. Results obtained by our models and the best performing models on each metric are included.

Model	1 post			100 posts			500 posts			1000 posts		
	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
TUA1#0	**0.80**	**0.88**	0.44	0.60	0.72	0.52	0.60	0.67	0.52	0.70	0.80	0.57
TUA1#1	0.70	0.77	0.44	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1#2	**0.80**	**0.88**	0.44	0.60	0.72	0.52	0.60	0.67	0.52	0.70	0.80	0.57
TUA1#3	0.60	0.69	0.43	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1#4	0.50	0.37	0.35	0.00	0.00	0.36	0.00	0.00	0.36	0.20	0.12	0.31
CYUT#3	0.10	0.07	0.12	0.70	0.70	0.57	0.70	0.72	0.59	**0.80**	0.74	0.60
CYUT#4	0.10	0.06	0.12	0.60	0.68	0.55	0.60	0.69	0.59	**0.80**	0.84	0.61
BLUE#0	**0.80**	**0.88**	**0.54**	0.60	0.56	0.59	0.80	0.81	0.66	**0.80**	0.80	0.68
BLUE#1	**0.80**	**0.88**	**0.54**	0.70	0.64	**0.67**	0.80	0.84	**0.74**	**0.80**	**0.86**	**0.72**
BLUE#2	**0.80**	0.75	0.46	0.40	0.40	0.30	0.30	0.35	0.20	0.30	0.38	0.16
NLPGroup-IISERB#0	0.00	0.00	0.02	**0.90**	0.92	0.30	**0.90**	**0.92**	0.33	0.00	0.00	0.00
NLPGroup-IISERB#1	0.30	0.32	0.13	**0.90**	0.81	0.27	0.80	0.84	0.33	0.00	0.00	0.00
NLPGroup-IISERB#4	0.00	0.00	0.04	**0.90**	0.93	0.66	**0.90**	**0.92**	0.69	0.00	0.00	0.00
UNED-MED#3	**0.80**	0.82	0.29	0.60	0.44	0.31	0.80	0.73	0.36	0.40	0.51	0.30
Sunday-Rocker2#1	0.70	0.81	0.39	**0.90**	**0.93**	0.66	**0.90**	0.88	0.65	0.00	0.00	0.00
Sunday-Rocker2#3	**0.80**	**0.88**	0.41	0.50	0.50	0.23	0.60	0.69	0.34	0.00	0.00	0.00
UNSL#1	**0.80**	**0.88**	0.46	0.60	0.73	0.64	0.60	0.73	0.66	0.60	0.71	0.66

indicator of depression and that a long affective memory is crutial for TAM to explore the users' affective states. Our initial experiment suggests that adding a latency penalty to the cross-entropy loss is effective for training early detection models. Among all participants, our system turns to be the most efficient and achieves two state-of-the-art results in terms of the ranking-based evaluation. Our results also suggest that language usage patterns, such as n-grams, could be an important feature for depression detection. Integrating language usage patterns into the TAM network could be a promising work in the future.

Acknowledgments

This research has been supported by JSPS KAKENHI Grant Number 19H04215.

References

[1] D. E. Losada, F. Crestani, J. Parapar, Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations., in: CLEF (Working Notes), 2017.
[2] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: early risk prediction on the

internet, in: International conference of the cross-language evaluation forum for european languages, Springer, 2018, pp. 343–361.
[3] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
[4] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) (2020).
[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview), CLEF (Working Notes) (2021).
[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Evaluation report of erisk 2022: Early risk prediction on the internet, CLEF (Working Notes) (2022).
[7] M. Park, D. McDonald, M. Cha, Perception differences between the depressed and non-depressed users in twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 7, 2013, pp. 476–485.
[8] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th annual ACM web science conference, 2013, pp. 47–56.
[9] J. Parapar, D. E. Losada, A. Barreiro, A learning-based approach for the identification of sexual predators in chat logs., in: CLEF (Online working notes/labs/workshop), volume 1178, 2012.
[10] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, T. Becker, Feeling bad on facebook: Depression disclosures by college students on a social networking site, Depression and anxiety 28 (2011) 447–455.
[11] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognition & Emotion 18 (2004) 1121–1133.
[12] S. J. Blatt, Experiences of depression: Theoretical, clinical, and research perspectives., American Psychological Association, 2004.
[13] M. Aaron T. Beck, P. Brad A. Alford, Depression: Causes and Treatment, University of Pennsylvania Press, 2014. URL: https://doi.org/10.9783/9780812290882. doi:doi:10.9783/9780812290882.
[14] J. Rottenberg, Mood and emotion in major depression, Current Directions in Psychological Science 14 (2005) 167–170.
[15] J. Joormann, C. H. Stanton, Examining emotion regulation in depression: A review and future directions, Behaviour research and therapy 86 (2016) 35–49.
[16] M. Stankevich, V. Isakov, D. Devyatkin, I. V. Smirnov, Feature engineering for depression detection in social media., in: ICPRAM, 2018, pp. 426–431.
[17] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T. S. Chua, W. Hall, Cross-domain depression detection via harvesting social media, International Joint Conferences on Artificial Intelligence, 2018.
[18] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from twitter activity, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp. 3187–3196.
[19] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, arXiv preprint arXiv:1709.01848 (2017).

[20] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893.
[21] A. Rinaldi, J. E. F. Tree, S. Chaturvedi, Predicting depression in screening interviews from latent categorization of interview prompts, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7–18.
[22] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, G. Long, Supervised learning for suicidal ideation detection in online user content, Complexity 2018 (2018).
[23] S. Ji, X. Li, Z. Huang, E. Cambria, Suicidal ideation and mental disorder detection with attentive relation networks, Neural Computing and Applications (2021) 1–11.
[24] L. Ansari, S. Ji, Q. Chen, E. Cambria, Ensemble hybrid learning methods for automated depression detection, IEEE Transactions on Computational Social Systems (2022).
[25] C. Yang, Y. Zhang, S. Muresan, Weakly-supervised methods for suicide risk assessment: Role of related domains, arXiv preprint arXiv:2106.02792 (2021).
[26] J. Joormann, I. H. Gotlib, Updating the contents of working memory in depression: interference from irrelevant negative material., Journal of abnormal psychology 117 (2008) 182.
[27] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.
[28] M. Trotzek, S. Koitka, C. M. Friedrich, Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression., in: CLEF (Working Notes), 2017.
[29] M. Trotzek, S. Koitka, C. M. Friedrich, Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia., in: CLEF (Working Notes), 2018.
[30] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y Gómez, M. Errecalde, Unsl's participation at erisk 2018 lab., in: CLEF (Working Notes), 2018.
[31] S. Paul, S. K. Jandhyala, T. Basu, Early detection of signs of anorexia and depression over social media using effective machine learning frameworks., in: CLEF (Working notes), 2018.
[32] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, Y.-S. Chen, CARER: Contextualized affect representations for emotion recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3687–3697. URL: https://www.aclweb.org/anthology/D18-1404. doi:10.18653/v1/D18-1404.
[33] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, J. Zhou, Patient subtyping via time-aware lstm networks, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 65–74.
[34] D. Zhang, J. Thadajarassiri, C. Sen, E. Rundensteiner, Time-aware transformer-based network for clinical notes series prediction, in: Machine Learning for Healthcare Conference, PMLR, 2020, pp. 566–588.
[35] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.