UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

A Time-Aware Mental State Space for Multimodal Depression Detection on Social Media

Permalink

https://escholarship.org/uc/item/04v1h1pq

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Nguyen, Quang Vinh
Nguyen, Dong Thanh
Nguyen, Duc Duy
et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Time-Aware Mental State Space for Multimodal Depression Detection on Social Media

Quang Vinh Nguyen
IT Center, Viettel Post, Viet Nam

Thanh Dong Nguyen
FPT Software Quy Nhon, Viet Nam

Duc Duy Nguyen
Hanoi University of Science and Technology, Viet Nam

Doan Khai Ta
Hanoi University of Science and Technology, Viet Nam

Hai Binh Nguyen
IT Center, Viettel Post, Viet Nam

Ji-eun Shin
Chonnam National University, South Korea

Seungwon Kim
Chonnam National University, South Korea

Hyung-Jeong Yang
Chonnam National University, South Korea

Soo-Hyung Kim
Chonnam National University, South Korea

Abstract

Detecting depression from user-generated posts on social media platforms offers significant potential for early intervention on at-risk individuals. Existing works mainly concentrate on text processing, and only a limited number incorporate images posted by users. These image-integrated methods face challenges in modeling the intricate relationships between textual and visual features. Besides, the absence of approaches that explore psychological trajectory of users by analyzing their posts over time leaves a critical gap in capturing the progression of depressive symptoms. In this paper, we propose A **T**ime-Aware **M**ental **S**tate **S**pace (**T-M2S**) for detecting depression from social media posts. We introduce a Cross-Modal Learning that effectively integrates text and image embeddings into sentiment-oriented unified representations. Additionally, we design a Mental State Space to analyze users' posts over time, offering a nuanced understanding of emotional dynamics. Extensive experiments on Twitter and Reddit datasets demonstrate that T-M2S significantly outperforms state-of-the-art methods. Code and models are available at GitHub.

Keywords: Depression detection; multimodal representation; social media

Introduction

The World Health Organization (WHO) (Organization et al., 2017) estimates that over 350 million people globally suffer from depression, and the incidence is increasing annually. Depression is a common mental disorder characterized by persistent feelings of sadness, hopelessness, and diminished interest in daily activities (Moustafa, Tindle, Frydecka, & Misiak, 2017). It significantly affects on individuals' emotional well-being, daily functioning, and quality of life. In severe cases, depression can lead to self-harm or even suicide (Huang et al., 2019). Early identification and intervention are critical to mitigating the long-term consequences of depression, such as impaired social relationships, reduced productivity, and an elevated risk of suicide. More than 70% of people in the early stages of depression would not consult the psychological doctors, deteriorating their conditions. On the other hand, social media platforms such as Twitter and Reddit (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017) often serve as a window into users' thoughts, emotions, and daily experiences, offering valuable insights into mental

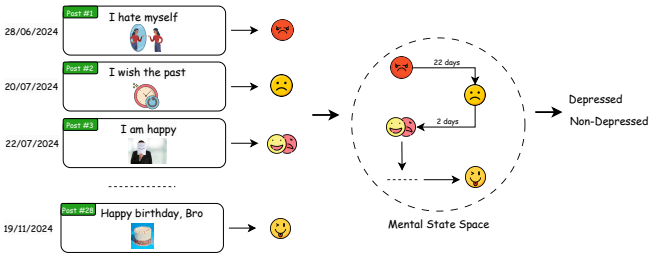


Figure 1: Mental State Space offers a detailed and nuanced understanding of psychological shifts and changes.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

health states. Harnessing the potential of social media posts for early depression detection presents a transformative approach to mental health care.

Existing studies (Alhuzali, Zhang, & Ananiadou, 2021; Kenton & Toutanova, 2019; Kim & Vossen, 2021) have primarily concentrated on the analysis of text-only social media posts, fuelled partly by the extensive availability of textual data and the remarkable advancements of pretrained language models (eg. BERT). These methods overlook the image modality which often carries contextual and emotional cues that textual content alone cannot capture (Guntuku, Preotiuc-Pietro, Eichstaedt, & Ungar, 2019; Reece & Danforth, 2017). Recently, multimodal approaches (Gui, Zhu, et al., 2019; Radford et al., 2021; Safa, Bayat, & Moghtader, 2022; Xu, Pérez-Rosas, & Mihalcea, 2020), that incorporate both textual and visual information have shown promising results. However, these methods often rely on simplistic fusion techniques, which are inadequate for modeling the intricate relationships and interactions between textual and visual features. Consequently, these methods struggle to fully leverage the complementary nature of these modalities, limiting their potential to provide accurate insights into users' mental health states. Temporal information is essential for understanding the progression of psychological states over time. Although some methods (Bucur, Cosma, Rosso, & Dinu, 2023; Sawhney, Joshi, Gandhi, & Shah, 2020; Zafar, Aftab, Qureshi, Wang, & Yan, 2024) incorporate temporal aspects, they often simplify the user timeline by treating it as a "bag-of-posts," where each post is analyzed independently. This approach disregards the chronological order and interdependence of posts, thereby failing to capture the underlying temporal dynamics and psychological trajectories. In summary, two main challenges remain: (1) multimodal approaches face struggle to effectively integrate textual and visual information into coherent, sentiment-oriented representations, and (2) the limited focus on temporal analysis to capture users' psychological trajectory over time. To address these, we propose **A Time-Aware Mental State Space**, a novel framework shown in Figure 1. We first introduce a Cross-Modal Learning to seamlessly integrates text and image embeddings to create unified, sentiment-oriented representations. Furthermore, we build a Mental State Space incorporating emotional cues extracted from text, and temporal information to analyze user posts over time and providing a nuanced understanding of psychological changes. Experiments on two benchmark datasets show our method outperforms state-of-the-art approaches.

In summary, our contributions are as follows:

- We propose **A Time-Aware Mental State Space (T-M2S)** for detecting depression from temporal, textual and visual information from social media posts.

- We introduce a **Cross-Modal Learning**, which effectively fuses text and image embeddings to generate sentiment-oriented representations. Additionally, we propose the **Mental State Space** to explore mood shift in user' posts over time, capturing the progression of psychological states.

- Our method achieves competitive performance against state-of-the-art models on benchmark datasets. We further provide in-depth analysis with rich empirical results to validate the effectiveness and significance of the proposed approach. Furthermore, we perform a qualitative analysis, which proves that our model is **interpretable** and allows for the selection of the most informative posts in a user's social media timeline.

Related Work

Depression Detection on Social Media

Early approaches (De Choudhury, Gamon, Counts, & Horvitz, 2013; Resnik et al., 2015; Burdisso, Errecalde, & Montes-y Gómez, 2019) in depression detection predominantly relied on traditional machine learning techniques, leveraging handcrafted features extracted from textual data. However, these methods required manual extraction of domain-specific features, which could be time-consuming and poor generalization in most of the unseen cases. With the advent of deep learning, methods such as CNNs (Rao, Zhang, Zhang, Cong, & Feng, 2020; Yates, Cohan, & Goharian, 2017), LSTMs (Trotzek, Koitka, & Friedrich, 2018; Skaik & Inkpen, 2020) and Transformer-based architectures (Bucur, Cosma, & Dinu, 2021; Wu & Qiu, 2021; Alhuzali et al., 2021) yielded promising results using only the textual information from users' posts. However, these methods overlook the image modality, which can offer critical contextual and emotional insights into a user's mental health. To address this, multimodal methods incorporating visual information emerged as promising alternatives (Mann, Paes, & Matsushima, 2020; Safa et al., 2022; Xu et al., 2020; Chiu, Lane, Koh, & Chen, 2021). For instance, (Shen et al., 2017) proposed a multimodal depressive dictionary learning framework that combines textual, visual, and behavioral information from Twitter users. (Kenton & Toutanova, 2019) introduced a CNN-based architecture that leverages Bidirectional Encoder Representations from Transformers (BERT) for textual features and CNNs for visual features, effectively integrating these modalities to improve depression detection. (Gui, Zhu, et al., 2019) employed a cooperative multi-agent reinforcement learning framework to select relevant textual and visual features, leading to improved accuracy. (Cong et al., 2018) utilized a Bi-LSTM with an attention mechanism to capture rich and complementary textual and visual representations. (An, Wang, Li, & Zhou, 2020) enhanced topic modeling by incorporating auxiliary tasks that focused on textual and visual features, further improving the performance of depression detection. However, methods largely neglect temporal cues, such as the chronological order or relative timing of posts, which can provide valuable context. There has

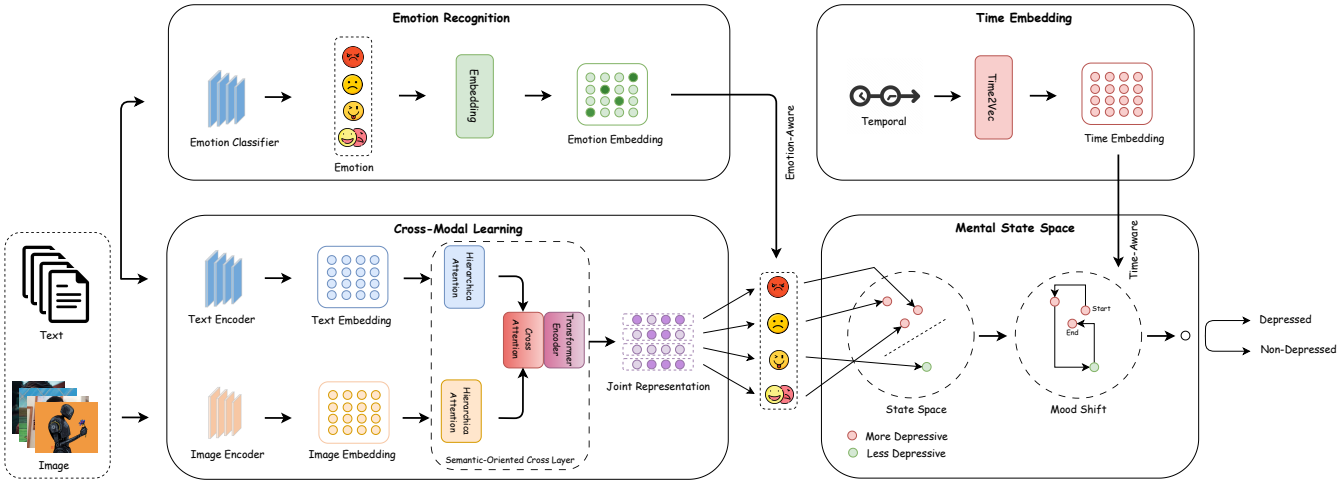


Figure 2: The overview of our framework. T-M2S is mainly divided into two stages: Cross-Modal Learning and Mental State Space. Cross-Modal Learning integrates textual and visual information to give join representations. Mental State Space captures the psychological changes in a user's posts over time.

been some recent works have begun addressing this limitation. (Sawhney et al., 2020) introduced a STATENet, a time-aware transformer that captures historical contexts of tweets. (Bucur et al., 2023) proposed a time-enriched multimodal transformer that incorporates temporal features alongside textual and visual representations for relative posting times. for relative posting time. (Zafar et al., 2024) developed a multimodal transformer that combines relative timestamps with image and text embeddings, further advancing time-aware multimodal depression detection.

Task Definition

Detecting Depression On Social Media is formulated as a user-level binary classification task. Given a user's social media timeline T, which is represented as:

$$T = [(t_1, T_1, V_1), (t_2, T_2, V_2), \ldots, (t_n, T_n, V_n)],$$

where t_i denotes the timestamp, T_i represents the textual content, and V_i corresponds to the visual content of the i-th post, the objective is to develop a model $F_\theta(\cdot)$ that predicts the binary label $y \in \{0,1\}$, where $y = 1$ indicates that the user is predicted to exhibit depressive symptoms and $y = 0$ indicates that the user is predicted to belong to the control group (non-depressed).

Proposed Framework

In this section, we describe our T-M2S framework. The framework is shown in Figure 2, which consists of two novel stages: Cross-Modal Learning and Mental State Space.

Cross-Modal Learning

Multimodal Input We consider a user to have multiple social media posts. We randomly sample L posts $\{P_i | i \in (1, L)\}$, where each post P_i consists of a text T_i and an accompanying image V_i. To extract meaningful embeddings from these modalities, we leverage state-of-the-art pretrained models. For the textual component, we utilize MentalBERT(Ji et al., 2022), a domain-specific transformer-based model designed for mental health-related text analysis. For the visual component, we employ DINO(Caron et al., 2021), a self-supervised vision transformer model. The projector ϕ is a learnable feed-forward layer that aligns the vision and text features by transforming the dimensions of the original representations to match the token dimension.

$$\mathbf{E}_V = \phi(\varphi_{\text{DINO}}(\mathbf{V})) \in \mathbb{R}^{L \times D} \quad (1)$$

$$\mathbf{E}_T = \phi(\varphi_{\text{MentalBERT}}(\mathbf{T})) \in \mathbb{R}^{L \times D} \quad (2)$$

We denote these sequence embeddings as $\mathbf{E}_V, \mathbf{E}_T \in \mathbb{R}^{L \times D}$, where L is the sequence length and D is the embedding dimension. For instance, on the Twitter(Shen et al., 2017) dataset, L varies in $\{128, 256, 512, 1024\}$ and D is set to 128.

Self Hierarchical Attention The standard attention mechanism(Vaswani et al., 2017) in transformers, while highly effective in many tasks, struggles to capture the distinct semantics in diverse social media posts. These posts often exhibit significant variations in sentiment, tone, and emotional context. To address this limitation, we propose an self hierarchical attention mechanism designed to acquire richer sentiment information relevant to depression detection. Instead of directly using $\mathbf{Q}$ and $\mathbf{K}$ vectors to interact with $\mathbf{V}$ as in prior works(Vaswani et al., 2017), we derive $\mathbf{Q}'$ and $\mathbf{V}'$ vectors from $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ thorough dot product and feed-forward neural network (FNN) operations allowing feature embeddings to better fit the task via multiple nonlinear projections. These deep vectors are then used to interact with the $\mathbf{V}$ vectors to weight important posts within

the input sequence. This operation allows the representation in each post to attend to all the representations in other posts, thereby inferring potential information and distinguishing posts that contain distinct sentiment information. Given visual and textual embeddings $\{\mathbf{E}_V, \mathbf{E}_T\}$, self hierarchical attention operations on each embedding can be formulated as following:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{E}W_Q, \mathbf{E}W_K, \mathbf{E}W_V \tag{3}$$

$$\mathbf{K}' = Norm(ReLU(FFN(\mathbf{K}))) \tag{4}$$

$$\mathbf{Q}' = Norm(ReLU(FFN(Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_q}}\right)\mathbf{V}))) \tag{5}$$

$$\mathbf{Z} = Softmax\left(\frac{\mathbf{Q}'\mathbf{K}'^\top}{\sqrt{d_q}}\right)\mathbf{V} \tag{6}$$

$$\mathbf{S}_T \leftarrow MHSA(\mathbf{Q}', \mathbf{K}', \mathbf{V}) \tag{7}$$

Where *MHSA* denotes multi-head self-attention operation(Vaswani et al., 2017), *FFN* refers to feed-forward neural network, *Norm* stands for layer normalization(Ba, Kiros, & Hinton, 2016), and *Softmax* represents weight normalization operation. $\mathbf{W} \in \mathbb{R}^{D \times D_k}$ are learnable parameters, D_k is the dimension of each attention head. In practice, we used 8-head attention and set D_k to 16.

Cross Attention After applying self attention to extract rich intra-modality features $\mathbf{S}_V, \mathbf{S}_T$ for each modality, we employ a cross-attention mechanism to facilitate the exchange of information between text and image modalities. This mechanism enables the model to learn complementary and joint cross-modality representations by aligning semantic entities across the two modalities. Given uni-modality representations $\{\mathbf{S}_V, \mathbf{S}_T\}$, we treat $\mathbf{S}_V$ as the query, and $\mathbf{S}_T$ as the key and value. Multi-head self-attention(Vaswani et al., 2017) is then conducted to obtain the cross-modality representations $\mathbf{C}$, as described below:

$$\mathbf{Q}_\mathrm{V}, \mathbf{K}_\mathrm{T}, \mathbf{V}_\mathrm{T} = \mathbf{S}_V W_Q, \mathbf{S}_T W_K, \mathbf{S}_T W_V \tag{8}$$

$$\mathbf{Z} = Softmax\left(\frac{\mathbf{Q}_\mathrm{V}\mathbf{K}_\mathrm{T}^\top}{\sqrt{d_q}}\right)\mathbf{V}_\mathrm{T} \tag{9}$$

$$\mathbf{C} \leftarrow MHSA(\mathbf{Q}_\mathrm{V}, \mathbf{K}_\mathrm{T}, \mathbf{V}_\mathrm{T}) \tag{10}$$

The terms in this formula correspond to those defined in the Self Hierarchical Attention subsection.

Transformer Encoder After obtaining the cross-modal representations $\mathbf{C}$ through the cross-attention mechanism, we apply a transformer encoder (Vaswani et al., 2017) to further refine these representations. This mechanism enables the encoder to capture long-range dependencies and intricate relationships within the integrated features from both modalities (text and image). This refined representation allows for a more coherent and enriched cross-modal understanding. Mathematically, given the cross-modal representation $\mathbf{C}$ as the input, the transformer encoder refines it as:

$$\mathbf{C} \leftarrow \mathrm{Transformer}(\mathbf{C}) \tag{11}$$

Mental State Space

State Space For each post, a pretrained emotion classifier (Hartmann, 2022) is used to predict textual content into one of seven emotion classes: anger, disgust, fear, joy, sadness, surprise, neutral. Given the text representation $\mathbf{T}$, the classifier outputs a one-hot encoded vector representing the predicted emotion. The predicted emotion is passed through an embedding layer to map it into a continuous vector space:

$$\mathbf{E}_e = \mathrm{Embedding}(\mathrm{Classifier}(\mathbf{T}) \in \mathbb{R}^{L \times D} \tag{12}$$

The cross-modality representation $\mathbf{C}$ is combined with the emotion embedding $\mathbf{E}_e$ to create the emotion-aware representation. This is achieved through concatenation followed by a fully connected layer to align the dimensions and fuse the information:

$$\mathbf{C} = \mathrm{FNN}([\mathbf{C}, \mathbf{E}_e]) \tag{13}$$

Mood Shift To incorporate temporal information into the state space, we use the Time2Vec encoding technique (Kazemi et al., 2019) to transform the timestamps of each post into a vector representation. Time2Vec is advantageous as it is invariant to time rescaling, avoids the need for hand-crafted time features, and is periodic and simple to implement. The temporal sequence $\mathbf{t} = [\mathbf{t_1}, \mathbf{t_2}, ..., \mathbf{t_L}]$ is encoded as:

$$\mathbf{t}' = \mathrm{Time2Vec}(\mathbf{t}) \in \mathbb{R}^{L \times D} \tag{14}$$

State Space Models (SSMs) are considered as linear time-invariant systems that map an input sequence $x(t)$ to a response $y(t)$ through a latent hidden state $h(t) \in \mathbb{R}^N$. These models are typically formulated with the state matrix including $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{N \times 1}$, and $C \in \mathbb{R}^{1 \times N}$, as follows:

$$h'(t) = Ah(t) + Bx(t) \tag{15}$$

$$y(t) = Ch(t) \tag{16}$$

Given the joint cross-modality representations $\mathbf{C} \in \mathbb{R}^{L \times D}$, where L represents the number of posts and D denotes feature dimension. We introduce the State Space Model (SSM) that achieves the input sequence $\mathbf{C}$ and transforms the input token sequence $\mathbf{X} = \{x_k\}_{k=1}^L$ of this sequence into the target token sequence $\mathbf{Y} = \{y_k\}_{k=1}^L$. The proposed SSM is characterized by three parameters: $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, ϕ , defined as follows:

$$h_k = \overline{\mathbf{A}} concat(h_{k-1}, x_k) \tag{17}$$

$$\phi_{selective}(h_k)[i] = \begin{cases} 1 \text{ if } \alpha(h_k[i]) \geq 0 \\ 0 \text{ if } \alpha(h_k[i]) < 0 \end{cases} \tag{18}$$

$$h_k \leftarrow h_k \times \phi_{selective}(h_k) \times \mathbf{t}'_k \tag{19}$$

$$y_k = \overline{\mathbf{B}} h_k \tag{20}$$

Where α stands for an activation function, with the *tanh* function used in our experiments.

For each element $\{x_k \in \mathbb{R}^{1\times D} | k \in [1, L]\}$ in the input sequence $\mathbf{C} \in \mathbb{R}^{L\times D}$, we first extend the state space x_k by concatenating it with an initialized hidden state $h_{k-1} \in \mathbb{R}^{1\times D}$. This concatenation enhances the state representation by incorporating information from the previous states. The extended states are then projected through a linear layer represented by the parameter matrix $\overline{\mathbf{A}} \in \mathbb{R}^{2D\times D}$ to align the state space. Next, we integrate a selection mechanism ϕ to further refine the representation. This mechanism allows the model to filter out irrelevant state information and retain pertinent state information. Temporal information is integrated into the process through a timestamp variable $\mathbf{t}'_k$, enriching the hidden state with temporal awareness and enabling the model to adapt to time-sensitive dynamics. Finally, the unified state space is obtained through another parameter matrix $\overline{\mathbf{B}} \in \mathbb{R}^{D\times D}$, resulting in the final state representation y_k. The SSM can be described as an autoregressive manner:

$$p(\mathbf{Y}|\mathbf{C}) = \prod_{k=1}^{L} p(y_k|\mathbf{C}, y_{<k}). \tag{21}$$

Experiments

Experimental Setup

Datasets We perform extensive experiments on two multimodal benchmark datasets. The Twitter(Gui, Zhu, et al., 2019) dataset contains 1,402 users diagnosed with depression and 1,402 control users. We adopt a five-fold cross-validation approach, consistent with the methodology used by (Gui, Zhu, et al., 2019) to ensure a fair comparison with prior work. The Reddit(Uban, Chulvi, & Rosso, 2022) includes 1,419 users from the depression class and 2,344 control users. Following the methodology of (Uban et al., 2022), the dataset is partitioned into subsets of 2,633 users for training, 379 users for validation, and 751 users for testing.

Baselines We compare T-M2S against several state-of-the-art methods, including *i.e.*, T-LSTM(Sawhney et al., 2020), LTSM+RL, CNN+RL(Gui, Zhang, et al., 2019), MTAL(An et al., 2020), MTAN(Cheng & Chen, 2022), GRU+VGG-Net+COMMA(Gui, Zhu, et al., 2019), GRU+VGG-Net+Unified advantages(Stanford, 2016), Co-Attention(Lu, Yang, Batra, & Parikh, 2016), Dual-Attention(Nam, Ha, & Kim, 2017), Modality-Attention(Moon, Neves, & Carvalho, 2018), SenseMood(Lin et al., 2020), EmoBERTa Transformer, Time2VecTransformer, Vanilla Transformer, SetTransformer(Bucur et al., 2023), MTEN(Zafar et al., 2024).

Experimental Results

We compare T-M2S against state-of-the-art approaches in the depression detection task on the Twitter and Reddit datasets, with results presented in Table 1 and Table 2, respectively. The model uses as input textual embeddings extracted from MentalBERT and visual embeddings extracted from DiNO. On the Twitter dataset, as shown in Table 1, T-M2S achieves outstanding performance compared to recent cutting-edge methods, as evidenced by F1, precision, recall, accuracy score. Specifically, T-M2S with a window size of 1024 achieves an F1 score of 0.986, a precision score of 0.982, a recall score of 0.989 and an accuracy score of 0.986, outperforming the second-best MTEN by 4.1%, 3.7%, 4.4%, and 4.2%, respectively. On the Reddit dataset, Table 2 shows that our method is superior to the other approaches. T-M2S with a window size of 1024 achieves an F1 score, precision, recall, accuracy score of 0.950, 0.977, 0.925, and 0.951, respectively, showing a 3.7%, 6.4%, 1.2%, and 2.5% improvement over MTEN. These results demonstrate that T-M2S effectively and accurately detects depression across different social media platforms.

Table 1: Comparison results on the Twitter dataset.

Method	Modality	F1	Precision	Recall	Accuracy
T-LSTM	T	0.848	0.896	0.804	0.855
EmoBERTa Transformer	T	0.864	0.843	0.887	0.861
LSTM + RL	T	0.871	0.872	0.870	0.870
CNN + RL	T	0.871	0.871	0.871	0.871
Co-Attention	T+I	0.865	0.871	0.863	0.866
Dual-Attention	T+I	0.848	0.848	0.848	0.848
Modality-Attention	T+I	0.864	0.868	0.862	0.866
SenseMood	T+I	0.936	0.903	0.870	0.884
MTAL	T+I	0.842	0.842	0.842	0.842
GRU + VGG-Net + Unified Advantages	T+I	0.865	0.866	0.865	0.866
GRU + VGG-Net + COMMA	T+I	0.900	0.900	0.901	0.900
MTAN	T+I	0.908	0.885	0.931	-
Vanilla Transformer	T+I	0.886	0.868	0.905	0.883
SetTransformer	T+I	0.927	0.921	0.934	0.926
Time2VecTransformer	T+I	0.931	0.931	0.931	0.931
MTEN	T+I	0.945	0.945	0.945	0.945
T-M2S (Window Size 128)	T+I	0.955	0.970	0.940	0.955
T-M2S (Window Size 256)	T+I	0.966	0.975	0.957	0.966
T-M2S (Window Size 512)	T+I	0.978	**0.986**	0.971	0.978
T-M2S (Window Size 1024)	T+I	**0.986**	0.982	**0.989**	**0.986**

Table 2: Comparison results on the Reddit dataset.

Method	Modality	F1	Precision	Recall	Accuracy
T-LSTM	T	0.831	0.825	0.837	0.872
EmoBERTa Transformer	T	0.843	0.828	0.858	0.879
Co-Attention	T+I	0.846	0.906	0.793	0.855
Dual-Attention	T+I	0.835	0.830	0.839	0.834
reddit	T+I	-	-	-	0.663
Vanilla Transformer	T+I	0.837	0.827	0.848	0.876
SetTransformer	T+I	0.902	0.878	0.928	0.924
Time2VecTransformer	T+I	0.869	0.869	0.869	0.901
MTEN	T+I	0.913	0.913	0.913	0.926
T-M2S (Window Size 128)	T+I	0.925	0.924	0.924	0.924
T-M2S (Window Size 256)	T+I	0.938	0.952	0.925	0.939
T-M2S (Window Size 512)	T+I	0.945	0.930	**0.961**	0.944
T-M2S (Window Size 1024)	T+I	**0.950**	**0.977**	0.925	**0.951**

Ablation Studies

Effects of Different Components To validate the impact of individual components in T-M2S, we conduct an ablation study by removing each component from the overall model individually and present obtained results on Twitter and Reddit datasets in Table 3. First, removing the proposed Mental State Space (M2S) results in a noticeable drop in F1, Precision and Recall, highlighting its pivotal role. Additionally, we remove the Self Hierarchical Attention (SHA) from the Sentiment-Oriented Cross Layer (SCL), leaving only the



Figure 3: Visualization of the distribution of more depressive and less depressive posts.

Table 3: Effects of different components. The experiment uses a window size of 128. "w/o" denotes "without".

Method	Twitter			Reddit		
	F1	Precision	Recall	F1	Precision	Recall
T-M2S	**0.955**	**0.970**	**0.940**	**0.925**	**0.924**	**0.924**
w/o M2S	0.935	0.935	0.935	0.907	0.902	0.912
w/o SHA	0.940	0.931	0.952	0.905	0.900	0.910
w/o SCL	0.933	0.934	0.930	0.900	0.902	0.898

cross-attention and transformer encoder operation. This also leads to a decrease in F1, Precision and Recall, demonstrating that SHA is crucial to the effectiveness of the SCL component. Furthermore, replacing the SCL mechanism with a simple addition operation to fuse text and image embeddings causes a significant decline in performance, underscoring that the SCL strategy is effective in learning robust cross-modality representations.

Visualization

Visualization of Cross-Modality Representations To illustrate the effectiveness of our SCL in inferring potential information and distinguishing posts which contain distinct sentiment information, we use PCA (Maćkiewicz & Ratajczak, 1993) to visualize the cross-modality representations of a specific user on Twitter diagnosed with depression in Figure 3. The red-marked posts indicate higher levels of depressive content, while those marked in green indicate lower levels. As shown, the red-marked and green-marked posts form distinct clusters. This observation suggests that SCL can effectively give sentiment-oriented representations and differentiate between more depressive and less depressive posts. Additionally, we sample several example posts from each category to further highlight this distinction.

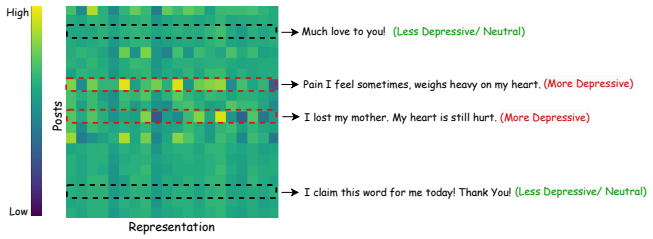


Figure 4: Visualization of the attention weights on user posts.

Time	Text	Emotion Cue
05:21:25 - 21/12/2016	**#1**, Normal day do some fun things over break, bake, clean	Joy
05:22:08 - 21/12/2016	**#2**, I'm exhausted I am one man doing the jobs of many	Sadness
14:02:04 - 21/12/2016	**#3**, I love having my morning snuggles with rescuedog	Joy
...	...	...
10:18:40 - 08/01/2017	**#30**, I hate when I'm wide awake at crazy hours then I blow hours on Pinterest looking at stupid stuff	Anger
14:57:56 - 09/01/2017	**#31**, Alright fans someone needs to take one for the team, go kidnap and drop him off	Neutral
06:52:31 - 10/01/2017	**#32**, Need to make something for dinner starving looks like it will be either eggs or a salad	Neutral
...	...	...
00:27:50 - 11/01/2017	**#59**, Left scene of 5-year-old death enter nearby apt complex	Sadness
02:18:29 - 11/01/2017	**#60**, Scene at apartment complex believed to be home to 5-year-old's relatives	Disgust
02:22:25 - 11/01/2017	**#61**, The gruesome detail of what police say lead to the death of that fice year old	Disgust

Label: Depressed **Ours:** Depressed **Ours (without M2S):** Non-Depressed **Co-Attention:** Non-Depressed
Dual-Attention: Non-Depressed **EmoBerta Transformer:** Non-Depressed **Set-Transformer:** Non-Depressed

Figure 5: Visualization of the mental state space of a depression user

Visualization of Post Attention In Figure 4, we present an attention map for posts of a depressed user on Reddit, showcasing the model's interpretability and explainability. As shown, our T-M2S assigns strongest attribution scores to posts with depression cues highlighting the posts that are more influential to the final prediction. This indicates that our method effectively filter and focus on relevant sentiment information from input to understand the user's mental state.

Visualization of Mental State Space We visualize the posts of a user from the Twitter dataset in Figure 5, which includes text, time, and emotional cues extracted from the user's posts. The visualization highlights the progression of the user's emotional state over time, with particular attention to the mood shifts indicated by the emotional tone in the text and the timing of the posts. Our T-M2S model, enhanced with the mental state space approach, demonstrates a strong ability to accurately detect signs of depression. The visualization further emphasizes the importance of considering the interplay between time, emotion, and text in accurately assessing mental health, showcasing the strength of our model in capturing these complex dynamics.

Conclusion

This paper proposes a novel approach called Time-Aware Mental State Space (T-M2S) for depression detection on social media. Our method effectively models multimodal data from users' posts and achieves competitive results on both Twitter and Reddit datasets. Through comprehensive analysis, we provide insights into the model's effectiveness, offering guidance for future research in mental health detection, particularly depression on social media.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (RS-2023-00256629) grant funded by the Korea government (MSIT), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00219107). The corresponding author is Soo-Hyung Kim.

References

Alhuzali, H., Zhang, T., & Ananiadou, S. (2021). Predicting sign of depression via using frozen pre-trained models and random forest classifier. In *Clef (working notes)* (pp. 888–896).

An, M., Wang, J., Li, S., & Zhou, G. (2020). Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics* (pp. 1078–1089).

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bucur, A.-M., Cosma, A., & Dinu, L. P. (2021). Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*.

Bucur, A.-M., Cosma, A., Rosso, P., & Dinu, L. P. (2023). It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European conference on information retrieval* (pp. 200–215).

Burdisso, S. G., Errecalde, M., & Montes-y Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, *133*, 182–197.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Cheng, J. C., & Chen, A. L. (2022). Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, *59*(2), 319–339.

Chiu, C. Y., Lane, H. Y., Koh, J. L., & Chen, A. L. (2021). Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, *56*(1), 25–47.

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., & Tao, C. (2018). Xa-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 1624–1627).

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the international aaai conference on web and social media* (Vol. 7, pp. 128–137).

Gui, T., Zhang, Q., Zhu, L., Zhou, X., Peng, M., & Huang, X. (2019). Depression detection on social media with reinforcement learning. In *Chinese computational linguistics: 18th china national conference, ccl 2019, kunming, china, october 18–20, 2019, proceedings 18* (pp. 613–624).

Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019). Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 110–117).

Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., & Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 236–246).

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Hartmann, J. (2022). *Emotion english distilroberta-base.* `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Huang, Y., Wang, Y., Wang, H., Liu, Z., Yu, X., Yan, J., ... others (2019). Prevalence of mental disorders in china: a cross-sectional epidemiological study. *The Lancet Psychiatry*, *6*(3), 211–224.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of lrec*.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., ... Brubaker, M. (2019). Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt* (Vol. 1).

Kim, T., & Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020). Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 407–411).

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, *29*.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (pca). *Computers & Geosciences*, *19*(3), 303–342.

Mann, P., Paes, A., & Matsushima, E. H. (2020). See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 440–451).

Moon, S., Neves, L., & Carvalho, V. (2018). Multimodal named entity recognition for short social media posts.

arXiv preprint arXiv:1802.07862.

Moustafa, A. A., Tindle, R., Frydecka, D., & Misiak, B. (2017). Impulsivity and its relationship with anxiety, depression and stress. *Comprehensive psychiatry*, *74*, 173–179.

Nam, H., Ha, J.-W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 299–307).

Organization, W. H., et al. (2017). Depression and other common mental disorders: global health estimates.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rao, G., Zhang, Y., Zhang, L., Cong, Q., & Feng, Z. (2020). Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, *8*, 32395–32403.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1), 15.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 99–107).

Safa, R., Bayat, P., & Moghtader, L. (2022). Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, *78*(4), 4709–4744.

Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. R. (2020, November). A time-aware transformer based model for suicide ideation detection on social media. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7685–7697). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.619 doi: 10.18653/v1/2020.emnlp-main.619

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... others (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Ijcai* (pp. 3838–3844).

Skaik, R., & Inkpen, D. (2020). Using twitter social media for depression detection in the canadian population. In *Proceedings of the 2020 3rd artificial intelligence and cloud computing conference* (pp. 109–114).

Stanford, M. E. (2016). Multi-agent deep reinforcement learning.. Retrieved from https://api.semanticscholar.org/CorpusID:265700928

Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, *32*(3), 588–601.

Uban, A., Chulvi, B., & Rosso, P. (2022, 09). Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality. In (p. 289-320). doi: 10.1007/978-3-031-04431-1_13

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wu, S.-H., & Qiu, Z.-J. (2021). A roberta-based model on measuring the severity of the signs of depression. In *Clef (working notes)* (pp. 1071–1080).

Xu, Z., Pérez-Rosas, V., & Mihalcea, R. (2020). Inferring social media users' mental health status from multimodal information. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 6292–6299).

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Zafar, A., Aftab, D., Qureshi, R., Wang, Y., & Yan, H. (2024). Multi-explainable temporalnet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2258–2265).