DynaProtect: A Dynamic Factor Influence Learning Framework for Protective Factor-aware Suicide Risk Prediction

Jun Li
The Hong Kong Polytechnic University
Hong Kong, Hong Kong
hialex.li@connect.polyu.hk

Xiangmeng Wang
The Hong Kong Polytechnic University
Hong Kong, Hong Kong
xiangmengpoly.wang@polyu.edu.hk

Yifei Yan
City University of Hong Kong
Hong Kong, Hong Kong
yfyan8-c@my.cityu.edu.hk

Haoyang Li
The Hong Kong Polytechnic University
Hong Kong, Hong Kong
haoyang-comp.li@polyu.edu.hk

Hong Va Leong
The Hong Kong Polytechnic University
Hong Kong, Hong Kong
cshleong@comp.polyu.edu.hk

Nancy Xiaonan Yu
City University of Hong Kong
Hong Kong, Hong Kong
nancy.yu@cityu.edu.hk

Qing Li*
The Hong Kong Polytechnic University
Hong Kong, Hong Kong
qing-prof.li@polyu.edu.hk

Abstract

Despite significant advances in approaches to suicide detection on social media, predicting users' suicide risk in a subsequent state remains challenging. Even though existing works have identified various risk factors to improve detection performance, they often overlook the critical role of protective factors in suicide prevention. To address this limitation, we propose an approach that jointly learns both risk and protective factors to predict users' subsequent suicide risk. Recognizing that the effectiveness of these factors varies across different user patterns, we introduce a dynamic factor influence learning mechanism that captures user-dependent interactions with risk and protective factors. Our experiments demonstrate that the integrated approach significantly enhances suicide risk prediction performance compared to existing methods.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Supervised learning*; • **Applied computing** → **Health informatics**.

Keywords

Suicide Risk Prediction, Suicide Prediction, Social Media, Big Data Processing, Risk Factors, Protective Factors.

ACM Reference Format:
Jun Li, Xiangmeng Wang, Yifei Yan, Haoyang Li, Hong Va Leong, Nancy Xiaonan Yu, and Qing Li. 2025. DynaProtect: A Dynamic Factor Influence Learning Framework for Protective Factor-aware Suicide Risk Prediction . In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3701716.3717510

*Corresponding author.

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717510

1 Introduction

Suicide is a global public health issue, resulting in approximately 726,000 deaths every year, particularly among young adults aged 15-29 [3]. Detecting at-risk individuals is critical for preventing life-threatening outcomes by proactive clinical and psychological interventions. Traditional approaches to detecting suicide thoughts often rely on expensive and time-consuming clinical procedures, such as questionnaires [6] and face-to-face consultation [25]. However, these approaches are limited in their reach, particularly for those with limited access to mental health resources or those hesitant to seek professional help.

Fortunately, social media platforms provide relatively anonymous and open spaces for individuals to discuss mental health issues. Recent advancements [13, 14, 22, 23] have utilized valuable data from social media to identify suicide risks, enabling more scalable and accessible methods for early detection and intervention.

Existing research on social media-based detection methods focuses mainly on incorporating additional user features and leveraging advanced model architectures. Early approaches primarily focused on identifying relevant user features to assess risk factors, such as psychological lexicons (e.g., LIWC [16]) and fundamental linguistic attributes like n-grams and POS tags.

For example, [24] extract LIWC features, statistical features, and POS tag counts from tweets as the source of risk factors and employs logistic regression and ensemble classifiers to identify suicide ideation. Due to the advancement of deep learning, recent work has focused on using various deep learning methods to detect risk factors in posts. For example, [11] proposed a multi-task learning framework to predict the future suicidal behavior of patients by

learning suicide risk levels together with their symptoms. Despite the efforts, these existing works suffer from critical limitations: (1) they mainly focus on modeling risk factors in user posts while neglecting the critical protective factors that provide emotional support and help users recover from suicidal tendencies. (2) they fail to capture the dynamic nature of suicidal risk in a subsequent state, where risk levels can vary significantly over short periods [19].

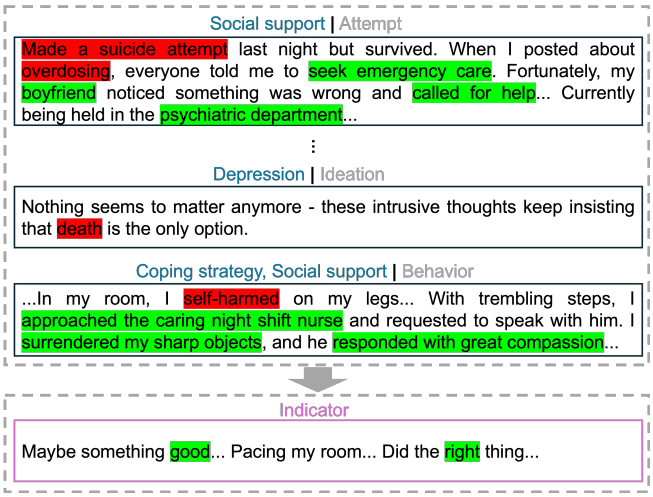

Figure 1: A toy example showing how both risk and protective factors are important for predicting a user's suicide risk in subsequent state, where red and green highlights indicate risk and protective expressions.

Protective factors, e.g., social support, coping strategies, and psychological capital, are equally important as risk factors in determining an individual's future suicide risk level. We use a toy example in Figure 1 to show the importance of protective factors. The social support for a suicide attempt user makes the user survive at first (suicide risk: from Attempt to Ideation). Similarly, in the second to last post, coping strategies de-escalating the user's suicide risk from Suicidal Behavior strong psychological capital. Additionally, predicting subsequent suicide risk is essential because suicide risk is inherently dynamic and time-sensitive. Traditional assessments fail to capture the temporal evolution nature of risk levels, whereas monitoring the interaction between different factors and subsequent suicide risk levels enables more timely and effective interventions.

Our work aims to predict suicide risk in a subsequent state given a user's history illustrated in Figure 2. Due to the distinct distributions of protective and risk factors in post sequences, we design two separate modules for learning these factors, avoiding potential learning bias. To predict the suicide risk of a subsequent state, we first identify critical factors influencing users' current states; we then develop an alignment learning framework that measures the associations between a user state timeline and potential factors. Furthermore, considering that the impact of both risk and protective factors on future suicide risk varies dynamically across different user state timelines, we propose a Bernoulli function to model factor activation patterns coupled with a novel loss function

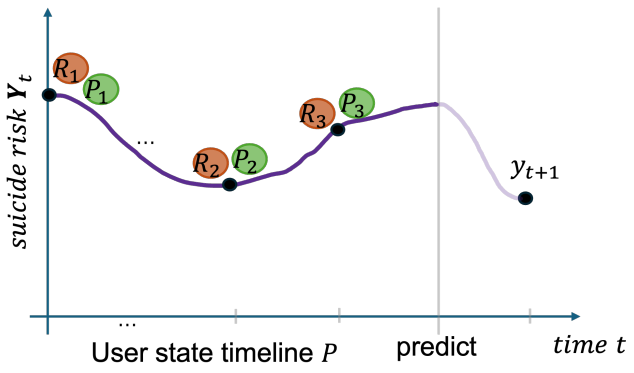

Figure 2: Given a user's historical posts, we predict their future suicide risk level y_{t+1}, where R and P denote risk and protective factors respectively.

to capture their time-varying influences on subsequent risk levels. In this study, we summarize our main contributions as follows:

- To the best of our knowledge, we are the first to integrate risk and protective factors modeling within a dynamic learning framework to predict subsequent suicide risk.
- We propose a novel alignment learning method that quantifies the impact of various factors on the current user state timeline, achieving a dynamic learning purpose.
- Extensive experiments show the superiority of our proposed method in the suicide risk prediction task, highlighting the importance of incorporating both risk and protective factors.

2 Related Work

Research on social media-based suicide detection has employed diverse features in predictive models. [27] demonstrated the significant role of emoji usage patterns in suicide ideation classification. UoS [1] developed an integrated approach that jointly models mood transitions and suicide risk through a pre-trained language model and attention mechanisms, effectively capturing contextual information across multiple posts. [13] incorporated suicide risk factors, while [12] joint learned bipolar disorder users' symptoms to enhance model sensitivity to critical indicators in posts. However, these approaches focus predominantly on various risk-related features while overlooking the potential impact of protective factors. This limitation emphasizes the need for a more comprehensive framework that incorporates both risk and protective factors in the prediction of suicide risk.

Recent studies have proposed various approaches to address different challenges in suicide risk assessment. SISMO [21] combines Longformer [2], bidirectional LSTM, and attention mechanisms to encode posting sequences for user suicide risk classification. STATENet [20] assesses post-level suicide risk by considering contextual information through a dual-branch architecture that jointly learns the current embedding of post and historical posts' representations. TSAML [12] implements a multi-task learning framework to predict users' maximum suicide risk over extended periods. However, suicide risks of users can fluctuate rapidly in the short term.

Its long-term prediction may potentially compromise intervention timeliness. As a result, existing approaches do not address the prediction of suicide risk in a user's subsequent state.

3 Protective Factor-aware Data Collection and Preprocessing

We collected posts from the 'r/SuicideWatch' subreddit of Reddit, following similar data collection methods used in previous Reddit-based studies [13]. Particularly, 6943 users from 15/06/2010 to 18/09/2022 were collected. Prior to annotation, we performed data preprocessing steps, including anonymization. We removed all personally identifiable information (e.g., names, addresses, emails, and links) to protect user privacy. To effectively track the users' suicide risk move over time, we extract the 237 users with 2515 posts who publish at least seven posts during a week on the 'r/SuicideWatch' subreddit. Three well-trained PhD students majoring in Psychology and Computer Science label the dataset. In the annotation process, we focused on two main types of labels: (i) Suicide factors (e.g., mental health disorder and substance use) and (ii) levels of suicide risk (Indicator, Ideation, Behavior, Attempt). In cases of inter-annotator disagreement, we conduct collaborative discussions among annotators to achieve consensus.

We build the suicide factors framework in two dimensions: (1) Risk factors including mental health issues, physical health characteristics, substance use, hopelessness, emotion dysregulation, low self-esteem, poor school performance, low socio-economic status, community-level interpersonal violence, prior self-harm or suicidal behavior, poor social support, interpersonal difficulty, dysfunctional family, exposure to others' suicide, stressful life events, traumatic experience, cognitive deficits, suicide means, and sexual orientation related issues; (2) Five protective factors: social support, coping strategy, psychological capital, sense of responsibility, and meaning in life.

For suicide risk annotation, based on the Columbia Suicide Severity Rating Scale (C-SSRS) [17], we labeled posts into four suicidality levels including indicator, ideation, behavior, and attempt following criteria established in [13].

4 Methodology

4.1 Task Formulation

Given a user's temporal sequence of posts, i.e., user state timeline, $P = \{p_1, p_2, ..., p_t\}$ over time steps $T = \{1, ..., t\}$, where the user $u \in U = \{u_1, u_2, ..., u_n\}$, our task is to predict their suicide risk at subsequent time step. Based on the C-SSRS assessment scale, our model will predict at-risk individuals into four hierarchical risk categories, from lowest to highest: Indicator (IN), Ideation (ID), Behavior (BR), and Attempt (AT). As a diathesis-stress framework, the Fluid Vulnerability Theory (FVT) characterizes suicide risk as a dynamic process that varies across individuals [19]. The task aims to predict the suicide risk of the future post p_{t+1} of u from $y \in \{IN, ID, BR, AT\}$. Figure 3 illustrates the overall architecture of our proposed model, which consists of three main components: (1) Post embedding: encode each post into a dense vector to capture semantic information. (2) Temporal context modeling: aggregate historical posts based on their relevance and temporal relationships and (3) Joint learning of temporal representations and factor interactions: model the complex interplay between protective/risk factors and temporal user state timeline to capture their mutual influence.

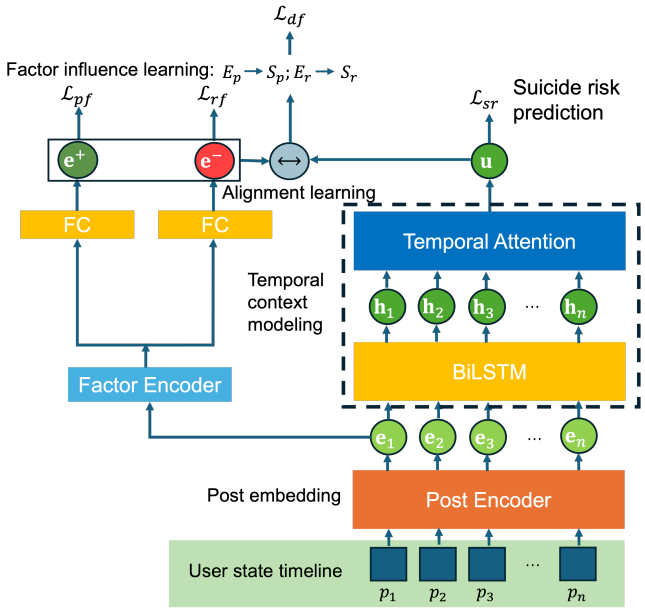

Figure 3: The overall architecture of the proposed model.

4.2 Post Embedding

Each post contains valuable user information that could indicate mental health states and suicide risk [4]. A sequence of posts can reveal the progression of a user's psychological states, which is crucial for suicide risk assessment [9]. To generate comprehensive semantic representations of posts, we employ Sentence-BERT (SBERT) [18], an extension of the pre-trained BERT model that has demonstrated effectiveness in representing user posts [11]. Formally, we denote the post embedding of each post p_t as $\mathbf{e}_t = SBERT(p_t)$.

4.3 Temporal Context Modeling

To address the limitation of point-in-time risk evaluations and capture the dynamic nature of risk indicators and their fluctuations over time [15], we propose to model the temporal evolution of user posts to reveal the critical patterns of mental state evolutions. Particularly, we first capture the sequential pattern of user's posts through *contextualize post modeling*. We then use *temporal attention* to assign different importance to the posts in the sequence.

4.3.1 Contextualize Post Modeling. We implement a BiLSTM [8] network that processes post embeddings bidirectionally, capturing temporal dependencies from both directions. We chose BiLSTM over Transformer [26] due to its computational efficiency and comparable performance in capturing local dependencies. Formally, the contextualized embedding of each post $\mathbf{e}_t$ is acquired by the following equation:

$$\begin{aligned}\overrightarrow{\mathbf{h}_t} &= \text{LSTM}(\mathbf{e}_t, \overrightarrow{\mathbf{h}}_{t-1})\\ \overleftarrow{\mathbf{h}_t} &= \text{LSTM}(\mathbf{e}_t, \overleftarrow{\mathbf{h}}_{t+1})\\ \mathbf{h}_t &= [\overrightarrow{\mathbf{h}_t}, \overleftarrow{\mathbf{h}_t}]\end{aligned} \tag{1}$$

4.3.2 Temporal Attention. While BiLSTM captures sequential dependencies, not all posts contribute equally to the user's current mental state. Recent posts may carry more weight in determining immediate risk, while certain historical posts containing critical emotion or significant behavioral changes may also be highly relevant. Therefore, we employ temporal attention [12] to dynamically weigh the importance of different posts across the user state timeline. We formulate the temporal attention mechanism as follows:

$$\mathbf{u} = \sum_{t=1}^{T} \mathbf{a}_t \mathbf{h}_t, \mathbf{a}_t = \frac{\exp(f(\boldsymbol{\delta}_t))}{\sum_{i=1}^{T} \exp(f(\boldsymbol{\delta}_t))}, \boldsymbol{\delta}_t = \sigma(\theta - \mu\Delta t)\mathbf{h}_t \tag{2}$$

where $\mathbf{u}$ is final attention output. $\mathbf{a}_t$ is temporal attention weight. $f(\cdot)$ is full-connected layer with tanh activation. $\boldsymbol{\delta}_t$ represents the hidden state incorporating temporal decay. Δt is time interval between posts. θ, μ is learnable parameters. σ is sigmoid function.

4.4 Joint Learning of Temporal Representations and Factor Interactions

We now present our joint learning strategy that captures the impact of protective and risk factors on dynamic user states upon model optimization.

4.4.1 Suicide Factors Learning. The distinct distributions of risk and protective factors, coupled with their simultaneous occurrence as shown in Figure 1, can result in biased feature representations that fail to capture the user's suicide risk accurately. To enhance the model's discriminative power, we implement a fully connected layer as a factor encoder and then separately learn the risk and protective factors in users' posts. It provides more nuanced insights into the dynamic balance between risk and protective factors while allowing us to understand how protective factors (e.g., social support, coping strategies) may buffer against risk factors (e.g., hopelessness).

Specifically, we design two parallel classification modules for risk and protective factors recognition. For each module, we employ the following loss to handle the multi-label classification task:

For risk factors:

$$L_{rf} = -\sum_{j=1}^{C_{rf}} [y_j^{rf} \log(\hat{y}_j^{rf})) + (1 - y_j^{rf}) \log(1 - \hat{y}_j^{rf})] \tag{3}$$

For protective factors:

$$L_{pf} = -\sum_{j=1}^{C_{pf}} [y_j^{pf} \log(\hat{y}_j^{pf}) + (1 - y_j^{pf}) \log(1 - \hat{y}_j^{pf})] \tag{4}$$

where C_{rf} and C_{pf} are the number of risk and protective factor labels respectively. y_j^{rf} and y_j^{pf} are the ground truth labels. $\hat{y}_j^{rf}$ and $\hat{y}_j^{pf}$ are the predicted logits.

4.4.2 Dynamic Factor Influence Learning. In future suicide risk assessment, the impact of protective and risk factors varies significantly depending on the context of the post sequence of different users. Notably, a single crucial protective factor can dominate and significantly reduce subsequent suicide risk, even in the presence of multiple risk factors. This dynamic interplay challenges traditional models that treat all factors with equal importance.

To address this challenge, we propose a dynamic factor influence approach that can identify and learn from dominant factors. Our approach consists of three key components: (1) effectiveness measurement of factors, (2) factor-state alignment learning, and (3) dynamic factor integration based on factor effectiveness.

Effectiveness measurement. First, we define how to measure the effectiveness of different factors in influencing suicide risk transitions. Let $\Delta risk = y_{t+1} - y_t$ represent the change in suicide risk between consecutive time steps. We introduce Bernoulli-based functions to capture factor effectiveness.

For protective factors, as protective factors could lower the risk of users' suicide attempts in the future, we thus define the Bernoulli function of protective factors as:

$$E_p = \begin{cases} 1 & \text{if } \Delta\text{risk} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Similarly, we define the Bernoulli function to denote the probability that risk factors elevate suicide risk:

$$E_r = \begin{cases} 1 & \text{if } \Delta\text{risk} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Factor-state alignment learning. To identify which factors are most relevant to a user's current state, we develop an alignment learning framework that measures the association between user state timeline and potential influencing factors. As shown in Figure 3, for each user state timeline embedding $\mathbf{u}$ acquired from Eq. (2), we define the association strength through the following functions:

$$S_p = -log\frac{exp(sim(\mathbf{u}, proj_p)/\tau)}{exp(sim(\mathbf{u}, proj_p)/\tau) + exp(sim(\mathbf{u}, proj_p)/\tau)} \tag{7}$$

$$S_r = -log\frac{exp(sim(\mathbf{u}, proj_r)/\tau)}{exp(sim(\mathbf{u}, proj_r)/\tau) + exp(sim(\mathbf{u}, proj_r)/\tau)} \tag{8}$$

where $proj_p = W_p[\mathbf{u}; \mathbf{e}_p]$ and $proj_r = W_r[\mathbf{u}; \mathbf{e}_r]$ project the user state timeline $\mathbf{u}$ and factors $\mathbf{e}$ into a shared space to measure their relevance. Here, $sim(\cdot, \cdot)$ quantifies the alignment strength through cosine similarity, and τ controls the sensitivity of the alignment measurement.

Dynamic factor integration. To capture the dynamic nature of factor influence, we propose a dynamic weighting mechanism that integrates the effectiveness measures of factors:

$$\begin{aligned}L_{df} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{2}\big(&E_p \cdot S_p + (1 - E_p) \cdot (1 - S_p)\\ &+ E_r \cdot S_r + (1 - E_r) \cdot (1 - S_r)\big)\Big]\end{aligned} \tag{9}$$

where E_p, E_r, S_p, and S_r are acquired from Eq. (5), Eq. (6), Eq. (7) and Eq. (8). n is the number of samples. This integrated loss function

adaptively adjusts the learning emphasis based on factor effectiveness: when protective factors show effectiveness ($E_p = 1$), the model prioritizes learning their patterns, similarly for risk factors.

4.4.3 Suicide Risk Prediction. To evaluate future suicide risk levels, we consider the ordinal relationship between different risk categories. We implement the ordinal regression loss [5] as an objective function for suicide risk prediction. For each true label k^a in the ordered set {0 : indicator, 1 : ideation, 2 : behavior, 3 : attempt}, we first calculate the absolute distances between k^a and each possible label value $k^i \in \{0, 1, 2, 3\}$ to form a distance vector $\phi = \alpha|k^a - k^i|$, where α is a penalty parameter for wrong predictions. Then, we obtain the distance-based probability distribution through $y_{sr} = \text{softmax}(-\phi)$. Formally, the distance-based probability distribution of user suicide prediction is represented by:

$$y_j^{sr} = \frac{e^{-\phi(k^i,k^a)}}{\sum_{j=1}^{L} e^{-\phi(k^i,k^a)}} \tag{10}$$

Where L is the number of suicide risk levels.

Finally, we use cross-entropy loss for suicide risk prediction:

$$\mathcal{L}_{sr} = -\sum_{j=1}^{L} y_j^{sr} \log(\hat{y}_j^{sr}). \tag{11}$$

4.4.4 Joint Learning. Our approach addresses different tasks at different granularities: suicide factors recognition on post and suicide risk prediction on user state timeline. To effectively balance these tasks, we adopt uncertainty-weighted loss [10]. It automatically learns optimal task weights by considering each task's inherent uncertainty, leading to our final objective function:

$$\begin{aligned}\mathcal{L}_{total} = &\frac{1}{2\sigma_1^2}\mathcal{L}_{sr} + \frac{1}{2\sigma_2^2}\mathcal{L}_{pf} + \frac{1}{2\sigma_3^2}\mathcal{L}_{rf} + \\ &\frac{1}{2\sigma_4^2}\mathcal{L}_{df} + \log(\sigma_1\sigma_2\sigma_3\sigma_4)\end{aligned} \tag{12}$$

where σ_1,σ_2,σ_3,σ_4 are learnable parameters, and $\log(\sigma_1\sigma_2\sigma_3\sigma_4)$ serves as a normalization term.

5 Experiments

5.1 Baselines

We adopt the comparative models from [12] as our baselines. These baseline models approach user representation learning from different perspectives. Furthermore, motivated by the recent emergence of large language models (LLMs) and their remarkable performance across various domains, we also incorporate several state-of-the-art LLMs as comparison models for this task.

- SISMO [21]: SISMO leverages Longformer [2] to encode individual posts, followed by a bidirectional LSTM layer and an attention mechanism for sequential modeling.
- STATENet [20]: STATENet employs a dual-branch architecture that jointly learns representations from historical tweets and the target tweet for final classification.
- TSAML [12]: TSAML employs Sentence-BERT to encode individual posts, followed by a bidirectional LSTM and temporal attention mechanism to identify critical suicide-related symptoms.

For large language models (LLMs), we incorporate state-of-the-art models, including GPT-3.5, GPT-4, and Claude-3.5-sonnet, as zero-shot baselines. Additionally, we performed instruction fine-tuning on Llama 3.1-8B and Gemma 2-9B-it using LoRA-based SFT on our dataset to establish strong supervised baselines for our experiments.

5.2 Results

5.2.1 Evaluation Metrics. In suicide risk prediction, due to the order nature of suicide risk level, we adopt modified metrics for suicide risk assessment evaluation, following [7]. The metrics redefine False Negatives (FN) as the ratio of under-predicted risk levels ($k^p < k^a$) and False Positives (FP) as the ratio of over-predicted risk levels. Precision and recall are adapted into Graded Precision (G.P.) and Graded Recall (G.R.) to account for the ordinal nature of suicide risk levels.

$$FN = \frac{\sum_{i=1}^{N_T} I(k_i^a > k_i^p)}{N_T}, FP = \frac{\sum_{i=1}^{N_T} I(k_i^p > k_i^a)}{N_T}$$

where k_i^a, k_i^p is the actual risk level and predict suicide risk. N_T is the size of testset.

5.2.2 Model Performance. Table 1 presents the comparative results of our proposed model against baseline approaches in the suicide prediction task. Our model outperformed both LLM-based methods and state-of-the-art approaches, achieving the highest Graded F1-score among all baselines. Among the LLMs, we observed that they are prone to having a higher G.R. score but lower G.P., indicating that cases of potential moderate risk are more likely to be classified as high risk. This "Better safe than sorry" approach makes them less likely to miss high-risk users during prediction. However, the low G.P. score indicates that LLMs default to safe predictions without comprehensively considering the protective and risk factors.

The other state-of-the-art approaches demonstrate a more balanced performance between G.P. and G.R. compared to LLMs. While STATENet achieved relatively higher G.R. scores, its inability to capture temporal patterns in sequential posts limited its overall performance, resulting in a lower G.F-score. Building upon existing baselines that utilize temporal context modeling, our model achieves superior results by incorporating risk and protective factors while capturing their interactions with the temporal user state timeline.

Table 1: Performance comparison of different models using graded metrics. G.P., G.R., and G.F-score denote Graded Precision, Graded Recall, and Graded F1-score respectively.

Model	G.P.	G.R.	G.F-score
Finetuned Llama 3.1-8B	0.5198	0.5756	0.5463
Finetuned Gemma 2-9B-it	0.4947	0.4312	0.4608
GPT-3.5	0.1603	0.8094	0.2676
GPT-4	0.2408	0.7849	0.3686
Claude-3.5-sonnet	0.2411	**0.8646**	0.3770
SISMO	0.6849	0.5018	0.5416
STATENet	0.4550	0.6224	0.5247
TSAML	0.6364	0.5000	0.5600
Our model	**0.8019**	0.5346	**0.6415**

5.2.3 Ablation Study. We conducted an ablation study to evaluate the contribution of each component in our model. The results in Table 2 demonstrate the effectiveness of our model architecture.

First, removing the dynamic factors integration learning (DF) component leads to a significant drop in performance, resulting in a lower G.F-score of 0.5657. This decrease indicates that dynamic factors integration learning plays a crucial role in learning more effective representations contributed by discriminative factors. Second, when protective factors are eliminated, the model's performance decreases notably (G.P.=0.6018, G.R.=0.5037, G.F=0.5484). This substantial drop in G.P. means it is essential to incorporate protective factors for accurate risk assessment, particularly in preventing false high-risk classifications. These results validate the effectiveness of our model design, where each component contributes meaningfully to the overall performance.

Table 2: Ablation study.

Model	G.P.	G.R.	G.F-score
Our model	0.8019	0.5346	0.6415
w/o DF	0.6311	0.5126	0.5657
w/o protective factors	0.6018	0.5037	0.5484

6 Discussion

Ethical Considerations: Our research on suicide risk prediction shows significant ethical challenges that require careful consideration. First of all, we emphasize that our model's predictions should not be considered a replacement for clinical assessment or professional medical judgment.

Privacy protection is paramount in our research methodology. To protect user privacy, we implemented robust de-identification procedures, which automatically removed all personally identifiable information from the collected posts. Posts in Figure 1 are paraphrased to protect user privacy. In addition, we acknowledge the potential unintended consequences of model application, which necessitates careful deployment and integration with existing professional support systems.

Limitation: Despite engaging mental health professionals from diverse backgrounds for training and annotation processes with cross validation, the assessment of suicide risk on posts is inherently subjective in nature, as the interpretation of risk signals can vary significantly across different contexts and individual experiences. While our model achieves higher G.P., the observed lower G.R. indicates a critical limitation, as missing potential suicide risk cases could have serious consequences.

7 Conclusion and Future Work

In this paper, we presented DynaProtect, a novel framework for suicide risk prediction that integrates both risk and protective factors. In addition, we propose a dynamic factor influence learning that effectively captures how different factors impact suicide risk across varying user state timelines. Our experiments show that DynaProtect consistently outperformed state-of-the-art baselines and LLMs. This improvement over these approaches demonstrates the effectiveness of our integrated framework. Future work will explore system scalability and real-world deployment.

Acknowledgments

We gratefully acknowledge Zehang Lin and Da Ren for their valuable suggestions and insightful discussions throughout this work. We also appreciate the domain experts who participated in the data annotation process.

References

[1] Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. 213–218.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR* abs/2004.05150 (2020). arXiv:2004.05150 https://arxiv.org/abs/2004.05150

[3] Fernando Cabrera-Eraso, Manuela Torres Solano, Mariana Vasquez-Ponce, Maria Alejandra Lopez-Orozco, Carlos Tirado, Maria Kamila Arevalo, Laura Medina, Catalina Rodriguez, Valentina Rodriguez-Castellanos, and Lina Maria Gonzalez. 2024. Effect of community interventions for the prevention of suicide in adolescents and young adults: a scoping review. *medRxiv* (2024), 2024–10.

[4] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 128–137.

[5] Raul Diaz and Amit Marathe. 2019. Soft Labels for Ordinal Regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 4738–4747. doi:10.1109/CVPR.2019.00487

[6] King-wa Fu, Ka Y Liu, and Paul SF Yip. 2007. Predictive validity of the Chinese version of the Adult Suicidal Ideation Questionnaire: psychometric properties and its short version. *Psychological Assessment* 19, 4 (2007), 422.

[7] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit P. Sheth, Randy S. Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *WWW*. ACM, 514–525.

[8] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).

[9] Sheri L Johnson, Charles S Carver, and Jordan A Tharp. 2017. Suicidality in bipolar disorder: The role of emotion-triggered impulsivity. *Suicide and Life-Threatening Behavior* 47, 2 (2017), 177–192.

[10] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 7482–7491. doi:10.1109/CVPR.2018.00781

[11] Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Towards Suicide Prevention from Bipolar Disorder with Temporal Symptom-Aware Multitask Learning. In *KDD*. ACM, 4357–4369.

[12] Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Towards Suicide Prevention from Bipolar Disorder with Temporal Symptom-Aware Multitask Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4357–4369.

[13] Jun Li, Xinhong Chen, Zehang Lin, Kaiqi Yang, Hong Va Leong, Nancy Xiaonan Yu, and Qing Li. 2022. Suicide risk level prediction and suicide trigger detection: A benchmark dataset. *HKIE Transactions Hong Kong Institution of Engineers* 29, 4 (2022), 268–282.

[14] Jun Li, Zhihan Yan, Zehang Lin, Xingyun Liu, Hong Va Leong, Nancy Xiaonan Yu, and Qing Li. 2021. Suicide Ideation Detection on Social Media During COVID-19 via Adversarial and Multi-task Learning. In *APWeb/WAIM (1) (Lecture Notes in Computer Science, Vol. 12858)*. Springer, 140–145.

[15] John L Oliffe, John S Ogrodniczuk, Joan L Bottorff, Joy L Johnson, and Kristy Hoyak. 2012. "You feel like you can't live anymore": Suicide from the perspectives of Canadian men who experience depression. *Social science & medicine* 74, 4 (2012), 506–514.

[16] James W Pennebaker. 2001. Linguistic inquiry and word count: LIWC 2001.

[17] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry* 168, 12 (2011), 1266–1277.

[18] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[19] M David Rudd. 2006. Fluid vulnerability theory: A cognitive approach to understanding the process of acute and chronic suicide risk. (2006).

[20] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In *EMNLP (1)*. Association for Computational Linguistics, 7685–7697.

[21] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021. Towards Ordinal Suicide Ideation Detectionon Social Media. In *Proceedings of 14th ACM International Conference On Web Search And Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA.

[22] Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021. Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning. In *NAACL-HLT*. Association for Computational Linguistics, 2176–2190.

[23] Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Ratn Shah, and Raj Singh. 2018. Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning. In *WASSA@EMNLP*. Association for Computational Linguistics, 167–175.

[24] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*. 91–98.

[25] Stefan Scherer, John Pestian, and Louis-Philippe Morency. 2013. Investigating the speech characteristics of suicidal adolescents. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 709–713.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[27] Tianlin Zhang, Kailai Yang, Shaoxiong Ji, Boyang Liu, Qianqian Xie, and Sophia Ananiadou. 2024. SuicidEmoji: Derived Emoji Dataset and Tasks for Suicide-Related Social Content. In *SIGIR*. ACM, 1136–1141.