DOI: **10.37943/25DSSM6014**

Yntymak Abdrazakh
Master of Science, Department of Computer Engineering
yntymak.abdrazakh@gmail.com, orcid.org/0000-0002-7119-1217
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
Batyrkhan Omarov
PhD in Information and Communication Technologies, Associate Professor,
Faculty of Computer Technology and CyberSecurity
batyahan@gmail.com, orcid.org/0000-0002-8341-7113
International Information Technology University, Kazakhstan
Aigerim Baimakhanova
PhD in Information System, Senior Lecturer, Department of Computer Engineering
aygerim.baymakhanova@ayu.edu.kz, orcid.org/0000-0002-5364-0146
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
Arypzhan Aben
Master of Science, Department of Computer Engineering
arypzhan.aben@ayu.edu.kz, orcid.org/0000-0001-8534-3288
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
Aigerim Sultanali
Master of Science, Department of Computer Engineering
aigerim.sultanali@ayu.edu.kz, orcid.org/0009-0008-6948-7271
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan

A TIME-AWARE TEMPORAL BERT FRAMEWORK FOR LONGITUDINAL DETECTION OF DEPRESSIVE AND SUICIDE-RELATED RISK PATTERNS IN SOCIAL MEDIA

Abstract: In this paper, we introduce a time-aware deep learning model designed to identify and predict signs of depression and suicidal ideation across social networks. Standard static text classifiers typically analyze updates in isolation; however, our method tracks the long-term progression of a person's emotional state by merging contextual language embeddings with temporal encodings and specific psycholinguistic markers. We gathered our primary dataset from Twitter, Reddit, and Facebook, ensuring all user histories were strictly anonymized and organized chronologically. The study evaluates multiple neural network architectures, specifically Temporal BERT (Bidirectional Encoder Representations from Transformers), time-encoded BiLSTM (Bidirectional Long Short-Term Memory), and a temporal transformer utilizing positional features. Our experiments demonstrate that factoring in the chronological dimension substantially boosts classification accuracy, allowing for the earlier detection of declining mental health. The Temporal BERT model achieved the highest F1 score (harmonic mean of precision and recall) and AUC (Area Under the Receiver Operating Characteristic Curve) values on several datasets, outperforming both standard (static) BERT and basic recurrent models. Analysis of temporal trajectories also allowed us to identify clear clusters of user behavior: stable, improving, and deteriorating - this makes conclusions more interpretable and helps us understand personal emotional dynamics. The early-warning module was evaluated at 7-, 14-, and 21-day prediction horizons and showed that risk-related deterioration patterns could be identified in advance of the reference event. Across all evaluated horizons, Temporal BERT demonstrated the strongest Recall@k performance, meaning that it more consistently captured at-risk users among the top-ranked predictions.

This article emphasizes that depressive and suicide-related risk signals are often not evident in isolated posts but emerge through longitudinal behavioral patterns. The proposed approach may support earlier and more sensitive identification of elevated risk patterns in digital mental health monitoring settings. At the same time, such use requires strict ethical safeguards, rigorous anonymization, and human-

Copyright © 2026, Authors. This is an open access article under the Creative Commons CC BY-NC-ND license
Received: 08.12.2025 Accepted: 25.02.2026 Published: 30.03.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

in-the-loop oversight. Future research should extend the framework toward multimodal, multilingual, and socially contextualized modeling.

Keywords: Temporal BERT; time-aware transformers; longitudinal mental health monitoring; depression detection; suicide risk detection; early warning modeling; psycholinguistic features; user-level trajectory analysis.

Introduction

The volume of depression-related content across online platforms has grown considerably, acting as an indirect gauge of escalating emotional stress globally. Empirical observations highlight a distinct correlation: heightened digital engagement among adolescents and young adults frequently parallels an uptick in depressive symptoms [1]. Because these networks now form a core layer of daily communication, they simultaneously act as a venue for self-expression and a potential catalyst for psychological struggles [2].

However, many existing machine learning models search for depression "specifically"-classifying each post individually, without taking into account the user's history and temporal context [3]. These approaches can indeed recognize warning signs in a specific post, but they are poor at revealing how a person's condition changes over time. This is important because mental health problems typically develop gradually: small changes accumulate over days or weeks, and these are precisely the changes that need to be detected early.

A more promising approach is to analyze texts dynamically. New research shows that temporal linguistic patterns, such as mood swings, changes in posting frequency, and the appearance or disappearance of certain lexical markers, can signal a worsening or improving condition long before a formal diagnosis [4]. Therefore, temporal models make it possible not only to identify risk but also to track the "trajectory" of a person's psychoemotional state.

It is worth noting that the use of time-aware architectures, such as Temporal-BERT (Bidirectional Encoder Representations from Transformers) or time-based recurrent models, in our case, can significantly increase the sensitivity and practical value of early detection systems. Instead of evaluating individual messages, such a system can observe the emotional dynamics of a specific user and identify early warning signs, which is particularly important for timely support and intervention.

We hypothesize that the most predictive indicators of deterioration are not isolated negative expressions, but sustained temporal patterns, including shifts in emotional tone, increased self-referential language, posting irregularity, and the accumulation of psycholinguistic risk markers over time.

1.1. Problem statement

Deep learning models can be effective at identifying depressive content in individual posts, but many of them share an important limitation: they operate in a static, post-by-post manner. When each message is evaluated in isolation, the model cannot capture how a person's mood and language shift across time. As a result, these systems are less suitable for early risk detection and continuous monitoring, where trends and gradual change are often the main signal.

The first problem is how to understand whether a person is getting worse or better over time. This requires looking not at a single post, but at consistent changes in language: a gradual decline in mood, more monotonous speech, an increase in self-references ("I", "me"), and an increase in negative emotions and themes. Longitudinal studies show that deterioration can manifest itself several weeks before overt episodes [5], but most NLP (Natural Language Processing) systems still rarely take time into account.

The second problem is whether it is possible to detect risk early, before direct statements about depression or suicide are made. Early signals are often very subtle: changes in post frequency and overall emotional tone, and the emergence of individual psycholinguistic markers that precede a sharp deterioration [6,7]. However, reliable models that can sensitively track such small changes over time are still few.

A third challenge concerns how temporal social media data should be prepared and represented. Posting behavior is irregular: one user may publish many messages in a day and then disappear for a week, while another posts in short bursts across different platforms. To model such histories, sequences need

DOI: 110.37943/25DSSM6014
© Yntymak Abdrazakh, Batyrkhan Omarov,
Aigerim Baimakhanova, Arypzhan Aben,
Aigerim Sultanali

careful normalization and construction so that both short-term fluctuations and longer-term trends are preserved. Although temporal transformers and time-aware LSTMs (Long Short-Term Memory networks) have performed well in other sequence tasks [8,9], their application to mental-health monitoring remains relatively limited.

This highlights a key gap in the research: a complete temporal deep learning model that can simultaneously:

1. Track changes in emotional state over time;
2. Provide early warning signs of risk even before overt crisis statements are made;
3. Work reliably and scalably with uneven and "dirty" time-based data from social media.

1.2. Aim and Objectives

The goal of the study is to create and test a time-aware deep learning model that can track changes in social media users' moods (depressive and suicidal signs). Conventional models analyze each post individually. We analyze the chain of posts and how language and emotions change over time. This allows us to detect gradual deterioration earlier.

Research Objectives:

- Build a longitudinal dataset with enough posts per user to analyze changes over time;
- Derive text-based features, including psycholinguistic indicators and contextual embeddings that capture meaning and affect;
- Develop time-aware models (e.g., Temporal BERT or a time-augmented LSTM) that explicitly use timestamps;
- Construct user "trajectories" by tracking how language indicators shift across weeks and months, distinguishing stability, improvement, and deterioration;
- Evaluate whether the model can raise early risk signals before explicit depressive or suicidal statements appear.

This approach helps us not only identify depressive words, but also understand how the condition develops and what typically occurs before a sharp decline.

1.3. Research Novelty

The novelty of this work is that we move from analyzing a single message to analyzing dynamics. Time is important because the condition changes gradually. We combine the meaning of the text, emotional cues, and time information in a single model to better detect changes.

Analytical Framework

2.1. Overview of the Analytical Approach

The approach consists of three parts. First, we analyze the language and meaning of posts to identify overt and subtle signs of depression or suicidal ideation [10]. To do this, we use contextual representations of text (embeddings) from transformer models-these help us better understand what is being expressed in the post.

Second, we consider time and examine how emotional signals change over days, weeks, and months. We construct a chronology of each user's posts and look for fluctuations, consistent trends, and early signs of deterioration that static models typically miss.

Third, we add behavioral indicators of online activity: how often a person posts, the length of pauses between posts, the extent to which their emotional tone changes, and other similar indicators. Such signals are often linked to psychological state and help more accurately describe the dynamics.

Together, these three components allow us to move beyond simply "searching for depressive posts" to analyzing the user's emotional trajectories. This becomes the basis for further development of temporal deep learning models.

2.2. Related Work and Research Gap

This study examines our temporal model in the general context of computational work on mental health. Previously, CNNs (Convolutional Neural Networks), LSTMs, and conventional transformers were most commonly used, which essentially solve the problem by taking a single post and classifying it. These models

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

are good at detecting overtly depressive language, but they struggle to capture the temporal relationships between posts and long-term behavioral changes. Therefore, their capabilities are often insufficient for real-world monitoring (when “early detection” is important).

Previously, it was shown that classical machine learning methods can detect depressive posts in social media feeds [11]. There are also studies examining neurophysiological and behavioral indicators (e.g., activity levels) that may be associated with depression [12]. AI chatbots for mental health support are also being studied separately: this is promising, but it faces many methodological and ethical limitations [13]. However, many of these studies neglect the key to our task: they don't construct user-level temporal dynamics, meaning they don't model how mood changes along the trajectory of posts. This is precisely what motivated our model, which tracks changes over time [14].

NLP approaches that specifically consider time are now emerging: temporal transformers, hierarchical sequence models, and longitudinal sentiment analysis. Their general idea is that mental health signals are often more evident not in a single word or post, but in how things change over time. Therefore, in this study, we compare temporal models with static ones based on several criteria: accuracy, early detection capability, and interpretability of results [15].

We also take into account that different platforms differ significantly. Twitter, Reddit, and Facebook have different communication styles, rules, and user behavior. Because of this, signs of depressive states may appear differently [16]. Therefore, models must be able to generalize and adapt well. By testing time models across different platforms, we evaluate not only the quality of predictions but also how robust they are across different digital environments.

Methods and Materials

3.1. Data Collection

The research used data from three social networks: Twitter, Reddit and Facebook. Only public posts, available for viewing by any user, were included in the selection. Our selection criteria prioritized users who demonstrated consistent publishing habits over an extended duration, allowing us to accurately construct publication timelines and monitor long-term deviations.

We identified relevant material using keywords and phrases associated with depression, suicidal thoughts, hopelessness, and severe stress. To support reliable temporal analysis, we included only accounts with at least 30 posts and continuous activity for a minimum of three consecutive months. This level of coverage makes it possible to observe both short-term shifts and more stable, long-term trends.

Before the analysis of the provided data, they were completely anonymized: user names, identifiers, mentions, links, e-mail addresses, geotags, and any other information that could be used to identify a person were removed. Instead, neutral substitutes such as <USER> and <URL> were used. Rare personal details that could indirectly indicate the author were also excluded.

For supervised modeling, each user timeline was assigned one of two labels: at-risk or stable. The at-risk label was assigned when the user’s sequence contained persistent depressive language markers, repeated expressions of hopelessness or self-devaluation, and a sustained increase in negative-affect signals across multiple posts or time windows. The stable label was assigned to users whose timelines did not exhibit persistent deterioration patterns and remained linguistically neutral or emotionally consistent over time. To reduce subjectivity, labels were derived using aggregated annotation criteria applied at the user level rather than on isolated posts.

The final dataset was split at the user level to prevent leakage between posts from the same individual across training and evaluation subsets. Figure 1 illustrates the main stages of the proposed Temporal BERT framework for risk detection in social media.

DOI: 110.37943/25DSSM6014
© Yntymak Abdrazakh, Batyrkhan Omarov,
Aigerim Baimakhanova, Arypzhan Aben,
Aigerim Sultanali

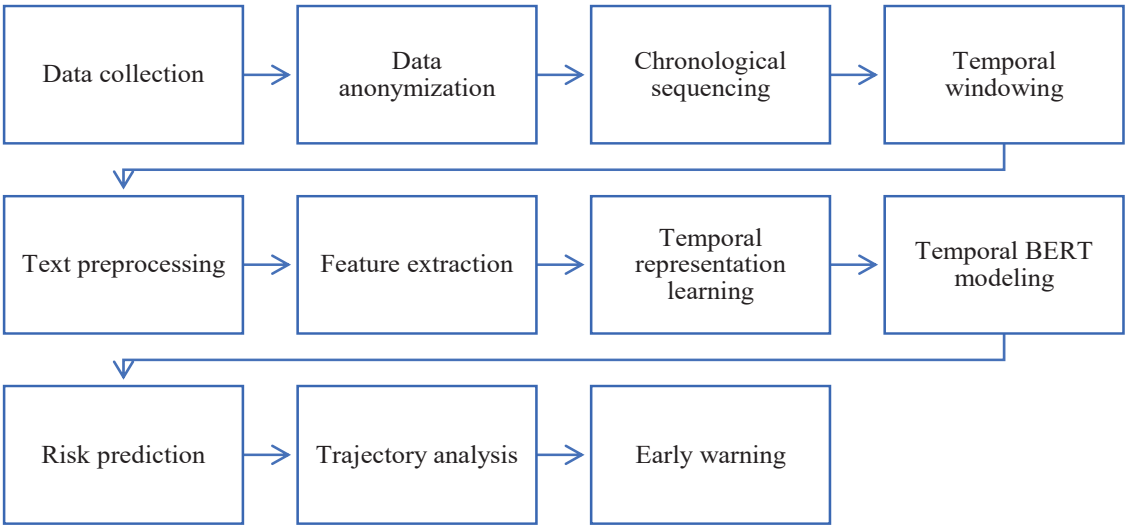

Figure 1. Block diagram of Temporal BERT for risk detection in social media

Only three parameters remained in the final set of data: text of publication, date/time of placement, and mark platform. Brief characteristics of the data set for each platform are given in Table 1.

Table 1. Summary statistics of the temporally structured dataset, including platform-wise user counts, post counts, and overall class distribution

Platform	Users	Posts	Avg. posts per user	Time span (months, median)	Avg. tokens per post
Twitter	120	5 480	45.7	3.4	18.2
Reddit	95	4 210	44.3	3.8	27.5
Facebook	85	3 960	46.6	3.2	22.9
Total	300	13 650	45.5	–	22.8

These statistics illustrate that the dataset is balanced in terms of temporal depth and posting frequency across platforms, providing a suitable basis for temporal modeling of depressive and suicidal language.

3.2. Preprocessing

We created a preprocessing pipeline to prepare raw social media posts for the temporal model without losing important linguistic cues associated with depression and suicide risk.

First, we standardized the texts: all content was lowercased, and links, user mentions, and platform-specific markup were removed. Hashtags were kept as separate tokens because they often carry meaning. Stop words were largely retained to avoid losing cues important for psycholinguistics, such as pronouns and negations (e.g., “not”, “never”). Emojis were replaced with simple emotion labels (e.g., “<EMO_HAPPY>”, “<EMO_SAD>”). Finally, we applied basic lemmatization to reduce word-form variation while preserving meaning.

Next, each user’s posts were sorted chronologically and grouped into fixed temporal windows. Daily windows were used for highly active users, whereas weekly windows were used for users with lower posting frequency to reduce sparsity and ensure comparability across timelines. Within each temporal window, we computed auxiliary activity features including the number of posts, the average token length, sentiment balance, and posting variability. By structuring the input in this manner, the model processes the textual semantics alongside dynamic shifts in user activity levels over time.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Alongside the text features, we added explicit temporal indicators. We measured gaps between posts (hours/days), overall posting frequency, and irregularity signals such as long pauses followed by sudden bursts. These temporal indicators were concatenated with contextual text representations to form a joint feature vector for each time step, allowing the models to capture both semantic content and temporal dynamics. Figure 2 shows the posts-per-user distribution; most accounts fall in the 30–60 range, which is sufficient for longitudinal analysis.

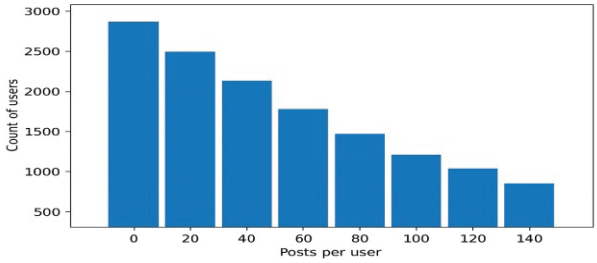

Figure 2. Posts per user in the collected dataset (all platforms). Most users contribute 30–60 posts, which provides enough temporal depth for longitudinal modeling.

3.3. Feature Engineering

To capture both semantic and psycholinguistic aspects of depressive language, we constructed a hybrid feature space that combines linguistic indicators with contextual and temporal embeddings. The feature groups are summarized in Table 2.

Linguistically, the system extracted lexicon-driven emotional markers by applying an NRC (National Research Council Canada) Emotion Lexicon-style affective dictionary alongside a LIWC (Linguistic Inquiry and Word Count)-inspired inventory of psychological categories. Across every temporal window and individual update, we calculated the normalized frequencies for positive and negative affective vocabulary, anxiety markers, first-person pronouns, negations, and high-risk expressions associated with self-harm or hopelessness (for example, “no reason to live” or “cut myself”). These variables provide interpretable signals that align with established accounts of depressive cognition.

At the representation level, each post was encoded using a pretrained BERT/RoBERTa (Robustly Optimized BERT Pretraining Approach) encoder, yielding dense contextual embeddings that preserve semantic dependencies beyond surface lexical counts [17]. For each user and each temporal window, post-level embeddings were aggregated using mean pooling and attention-weighted pooling, where higher weights were assigned to posts with stronger depressive probability estimates. To explicitly model time, the aggregated embeddings were concatenated with temporal encodings derived from normalized timestamps, including day-of-week, relative week index, and elapsed time since the first observed post. With this design, the model can differentiate earlier versus later stages in a user’s trajectory, even when users’ absolute dates do not align.

Formally, for each user u and time window t, the fused representation was defined as:

$$z_t = [e_t; \tau_t; a_t], \tag{1}$$

where e_t denotes the contextual text embedding, τ_t the temporal encoding, and a_t the activity-based auxiliary features.

DOI: 110.37943/25DSSM6014
© Yntymak Abdrazakh, Batyrkhan Omarov, Aigerim Baimakhanova, Arypzhan Aben, Aigerim Sultanali

Table 2. Main feature groups used in the experiments

Group	Description	Examples -Dimensions
Emotion lexicon	NRC-style counts and proportions of affective words	Neg/Pos emotion, anger, sadness (8 dims)
LIWC-inspired-like categories	Psycholinguistic categories linked to depression	Self-focus, social, cognitive processes (10)
Pronouns & negations	Normalized counts of singular self-references (e.g., “I”, “me”) combined with standard negations	1st-person sg., total negations (4)
Risk expressions	Dictionary of hopelessness and self-harm phrases	Binary flags + normalized counts (5)
Contextual embeddings	BERT / RoBERTa sentence embeddings per post/window	768-dimensional vectors
Temporal encodings	Sin/cos encoding of time indices and intervals	Week index, Δt between posts (8)
Activity indicators	Posting frequency, mean length, variability per window	#posts, avg tokens, variance (5)

3.4. Model Design

We compared three temporal architectures that differ in the way they integrate linguistic and temporal information, while keeping the input feature space identical.

For each user, the input was represented as an ordered sequence:

$$X_u = \{z_1, z_2, \dots, z_T\}, \quad (2)$$

where each vector z_T corresponds to one temporal window and combines semantic, temporal, and activity information. The modeling objective was to estimate either a post-level depressive label, a user-level risk trajectory, or an early-warning signal for future deterioration.

Temporal BERT.

In the Temporal BERT configuration, each fused window representation z_T was projected into a shared hidden space and processed by a lightweight transformer stack with additional time-aware positional encodings. This allowed the model to attend not only to semantic similarity between windows, but also to their relative temporal order and spacing.

A dedicated sequence summary token was used to obtain a user-level representation, from which the final depressive-risk score was predicted. In parallel, intermediate hidden states were used to estimate window-level risk values for trajectory reconstruction.

BiLSTM (Bidirectional Long Short-Term Memory) + Time Encoding.

In the BiLSTM + Time model, the fused vectors z_T were passed through a bidirectional LSTM to capture sequential dependencies in both forward and backward temporal directions. The hidden states were aggregated using attention pooling across time steps, yielding a user-level representation for downstream risk prediction. This architecture explicitly models forward and backward dependencies in the sequence but lacks the self-attention mechanism of transformers.

Temporal Transformer.

The Temporal Transformer model operated directly on the fused time-step vectors and incorporated sinusoidal temporal encodings at each sequence position, enabling the encoder to represent both local and long-range temporal dependencies. Unlike Temporal BERT, which relies on a pretrained language model only at the sentence level, the temporal transformer operates purely on the sequence of fused feature vectors and is trained from scratch for the temporal prediction task.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

All models were optimized with AdamW, using early stopping on a validation set. Hyperparameters were selected via a small grid search; the final configuration is shown in Table 3.

Table 3. Key hyperparameters of temporal models

Parameter	**Temporal BERT**	**BiLSTM + Time**	**Temporal Transformer**
Hidden size	256	256	256
# temporal layers	2	1 BiLSTM	3 encoder layers
Dropout	0.20	0.30	0.20
Learning rate	$2 \cdot 10^{-5}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$
Batch size (timelines)	16	32	32
Max sequence length	32 windows	32 windows	32 windows
Optimizer	AdamW	AdamW	AdamW

Results

4.1. Model Performance

To evaluate the contribution of temporal modeling, we included a regular BERT baseline that processes each post independently without modeling the temporal order of user activity. In contrast, Temporal BERT incorporates the sequence and timing of posts across the user history.

Incorporating temporal information improved predictive performance. Temporal BERT achieved an F1-score of 0.930, compared with 0.900 for Regular BERT, while AUC (Area Under the Receiver Operating Characteristic Curve) increased from 0.950 to 0.970. These results indicate that modeling the temporal evolution of user behavior provides a measurable advantage over isolated post-level classification.

ROC (Receiver Operating Characteristic) and Precision–Recall analyses further confirmed the advantage of Temporal BERT across decision thresholds. In particular, the temporal model showed better discrimination in settings where improved ranking of high-risk users is critical.

At Level 1 (post-level text classification), we used the same architectures to predict the depressive label of individual posts, treating each message as an independent input while retaining time encodings as auxiliary features. The main goal of this level was to ensure that temporal extensions do not degrade baseline classification performance. Table 4 presents the post-level results averaged over five user-level random splits, ensuring that posts from the same user did not appear in both training and evaluation subsets. Figure 3 shows a comparison of F1-score (harmonic mean of precision and recall) and AUC across the regular and temporal architectures

Table 4. Post-level classification performance of regular and temporal architectures averaged over five user-level random splits

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Temporal BERT	0.920	0.910	0.940	0.930	0.970
BiLSTM + Time	0.890	0.880	0.920	0.900	0.950
Temporal Transformer	0.910	0.900	0.930	0.920	0.960
Regular BERT	0.890	0.890	0.910	0.900	0.950

To quantify the contribution of temporal modeling, we included a Regular BERT baseline that processes posts independently without modeling the temporal order of user activity. In the post-level evaluation, the Regular BERT baseline achieved an Accuracy of 0.890, Precision of 0.890, Recall of 0.910, F1-score of 0.900, and AUC of 0.950. These results indicate that while Regular BERT provides strong post-level classification performance, incorporating temporal information yields a further improvement in discrimination quality.

DOI: 110.37943/25DSSM6014
© Yntymak Abdrazakh, Batyrkhan Omarov,
Aigerim Baimakhanova, Arypzhan Aben,
Aigerim Sultanali

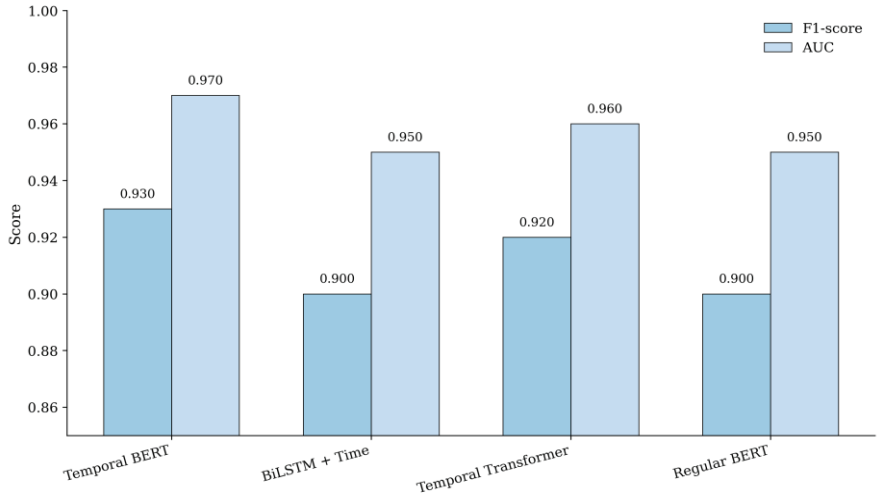

Figure 3. Comparison of F1-score and AUC across the regular and temporal architectures

For binary risk prediction, the output layer used a sigmoid activation, and the models were trained with binary cross-entropy loss. For user-level trajectory estimation, the predicted weekly scores were optimized to preserve both classification quality and temporal consistency. Early-warning performance was evaluated by checking whether the predicted risk exceeded the predefined alert threshold within the required lead-time window.

4.2. Evaluation Protocol

The evaluation protocol was structured in three levels to reflect different aspects of the temporal prediction task.

Level 1 – Text-level classification

At the first level, we evaluated standard post-level depressive classification using F1-score and AUC, as shown in Table 4. This stage verifies that temporal models maintain or improve upon the performance of non-temporal baselines.

Level 2 – User-level dynamics

At the second level, the goal was to assess how well the models approximate the trajectory of user-level depressive risk. For each user, weekly reference risk scores were derived using aggregated annotation or proxy supervision procedures, and these scores served as the temporal target for trajectory reconstruction. The models produced a continuous risk index R_t for each week t.

Two main metrics were used:

Correlation (r) between the predicted risk curve and the reference temporal signal;

RMSE (Root Mean Square Error) of the slope of the emotional curve, quantifying how accurately the model reproduces the direction and steepness of change.

A static BERT baseline was used for post-level comparison only. Since it does not explicitly reconstruct user-level temporal trajectories, it was not included in the trajectory-fit evaluation shown in Table 5.

Table 5. Level 2 – Fit to user-level dynamics

Model	Correlation r	Slope RMSE
Temporal BERT	0.780	0.110
BiLSTM + Time	0.710	0.150
Temporal Transformer	0.750	0.130

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Temporal BERT achieved the highest correlation with the reference trajectories and the lowest slope RMSE, indicating a better fit to the temporal evolution of emotional risk than the other temporal architectures.

Level 3 – Early-Warning Prediction

At the third level of assessment, we considered deterioration events as binary outcomes: whether the event occurred or did not occur. A deterioration event was defined as the earliest time point at which the reference risk trajectory crossed a predefined risk threshold or showed a sustained increase in depressive markers over consecutive windows.

For each user, we identified a reference time point t^*, defined as the first temporal window at which the reference risk trajectory crossed the predefined alert threshold. The goal was not to recognize deterioration after it happens, but to issue a warning beforehand: no later than H days before t^*. If the system raised an alert before this point, it indicates prospective prediction rather than retrospective detection [18]. We evaluated early-warning performance with two metrics:

Early warning accuracy representing the proportion of deterioration events correctly predicted within the allowed temporal window.

Recall@k for high-risk users, defined as the proportion of users whose deterioration was successfully detected among the top *k* percent ranked by predicted risk before the worsening occurred. This metric reflects the model's ability to prioritize users at greatest risk.

All evaluation procedures were performed at the user level to preserve the longitudinal structure of the data and avoid leakage across temporal windows.

Table 6 presents the early-warning results at prediction horizons of 7, 14, and 21 days.

Table 6. Level 3 - Early-warning performance at different prediction horizons

Horizon (days)	Model	Early warning accuracy	Recall@k (top 10%)
7	Temporal BERT	0.84	0.86
	BiLSTM + Time	0.79	0.80
	Temporal Transformer	0.82	0.83
14	Temporal BERT	0.80	0.81
	BiLSTM + Time	0.74	0.75
	Temporal Transformer	0.77	0.78
21	Temporal BERT	0.76	0.77
	BiLSTM + Time	0.70	0.70
	Temporal Transformer	0.73	0.73

Reported post-level results were averaged across five user-level random splits. To reflect variability across runs, the mean values should be interpreted together with the corresponding standard deviations.

The experiments were implemented in Python using PyTorch and the HuggingFace Transformers library. Auxiliary preprocessing and evaluation routines were performed with standard scientific computing libraries. Model training used AdamW optimization with early stopping on the validation subset. Hyperparameters were selected by a limited grid search over learning rate, dropout, batch size, and sequence depth. All experiments were run under the same training protocol to ensure comparability across models.

4.3. Temporal Trajectory Analysis

A key objective of the study was to examine whether temporal models could accurately reconstruct emotional trajectories over long observation periods. For each user, weekly aggregated risk scores were plotted to visualize the dynamic evolution of depressive indicators.

Temporal BERT captured three broad trajectory types – stable, improving, and worsening emotional patterns. Stable users showed only a small variation in predicted risk and relatively consistent language. Improving trajectories were characterized by a gradual reduction in depressive expression across several

DOI: 110.37943/25DSSM6014
© Yntymak Abdrazakh, Batyrkhan Omarov,
Aigerim Baimakhanova, Arypzhan Aben,
Aigerim Sultanali

weeks. Worsening trajectories, in contrast, exhibited steady increases in predicted risk, often together with higher variability in emotional markers and more frequent negative-affect vocabulary.

Clustering the trajectories using k-means (k = 3) revealed that there are indeed three main behavior patterns. Within each cluster, Temporal BERT's trajectories were the most similar. This means that its predictions are better at "aligning" with understandable and explicable changes in emotion than BiLSTM or Temporal Transformer.

Inspection of example trajectories suggests that deterioration often starts gradually rather than abruptly. Early changes can be subtle and easy to miss in a single post. For instance, a user may begin using first-person references ("I/me") more frequently and rely more on negative constructions (e.g., "not", "never"). Such shifts may appear weeks before depressive thoughts become explicit in the text. This observation helps explain why temporal models can pick up early linguistic cues that static models often overlook.

4.4. Early Warning Capability

To make the lead-time analysis more explicit, the early-warning results should be interpreted as horizon-specific prediction performance. In our setting, the model was evaluated at 7-, 14-, and 21-day horizons, which makes it possible to compare how predictive utility changes as the forecast window becomes longer. This formulation provides a clearer interpretation of early warning than a single aggregate statement about advance detection.

Temporal BERT achieved the strongest early-warning performance among the evaluated temporal models. At the 7-day horizon, Temporal BERT achieved a Recall@k of 0.86 and an early-warning accuracy of 0.84. Performance remained comparatively robust at the 14-day horizon (Recall@k = 0.81), with a gradual decline by 21 days.

Inspection of prediction sequences highlighted several recurring patterns that tend to precede deterioration:

- increasing focus on hopelessness and personal inadequacy in language;
- abrupt increases in posting frequency after periods of inactivity;
- semantic drift toward more negatively valenced vocabulary.

Across anonymized timelines, elevated risk often became visible days before overt expressions of emotional crisis. Taken together, these patterns underline the practical value of temporal models for monitoring and intervention settings.

4.5. Case Studies

To demonstrate how the framework can be interpreted in practice, we report several anonymized case studies drawn from user-level trajectories.

Case 1: Gradual Deterioration

The user began with a low and relatively stable predicted risk. Over the next six weeks, the model observed a rise in negative emotional language, more frequent use of first-person references ("I/me"), and a shift in content toward more depressive themes [19]. The risk curve started increasing about 12 days before the clinical threshold was crossed, enabling an advance warning rather than a retrospective flag.

Case 2: Rapid Risk Increase Following a Posting Gap

This user posted irregularly and had prolonged periods of silence. Temporal BERT detected a sharp jump in negative sentiment immediately after a 10-day break and issued an early-warning signal 7 days before the escalation became explicit in the language.

Case 3: Sustained Reduction in Risk Indicators

The user was initially classified as high risk but gradually moved toward more neutral linguistic patterns. A decrease in negations, increased use of collective pronouns ("we", "us"), and more stable sentiment coincided with declining risk scores. Overall, the trajectory was consistent with emotional improvement and illustrates the model's sensitivity to positive change.

Case 4: Episodic Variability in the Risk Trajectory

Predicted risk for this user alternated between peaks and valleys, mirroring episodic depressive language. Temporal patterns, together with periodic surges in emotional intensity, helped the model separate short-lived fluctuations from sustained deterioration.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

The trajectory graphs demonstrate that temporal modeling provides a more nuanced understanding of each individual's state: it helps not only identify risks but also explain why the model concluded they did.

Discussion

5.1. Interpretation of Results

The results showed that Temporal BERT almost always outperforms conventional static models and other temporal models across all assessment stages. This is because it simultaneously considers the meaning of the text (via BERT embeddings) and time (via temporal features). Therefore, the model sees not only the words but also how a person's state changes over time. Unlike conventional BERT, which treats each post as separate, Temporal BERT views a series of posts as a single story. This allows it to detect small changes that often appear before noticeable deterioration.

The analysis also revealed that certain linguistic features are often associated with deterioration. These include, for example, an increase in the use of "I/me" words, more negative expressions ("not", "never"), phrases about hopelessness, and more "heavy" emotional vocabulary. Behavioral changes are also important: for example, a person may be silent for a long time and then suddenly start posting more frequently. These features also help predict risk. Ultimately, it's clear that deterioration has its own "traces" in time and language, and temporal models are adept at capturing them.

5.2. Strengths of the Approach

The main advantage is that the model analyzes dynamics, not a single post. Therefore, it identifies risks that static models often miss. It is sensitive to both sudden episodes and gradual declines that stretch over weeks.

A second advantage is personalization. Because the model observes a user over time, it can learn the individual's typical writing style and flag departures from that baseline. This matters in practice: people differ widely, and the same words can carry different implications depending on the user and context.

Third, the framework supports early warning. Detecting elevated risk 7–14-21 days before deterioration provides a window for response, including follow-up, specialist attention, closer monitoring, or other preventive actions.

5.3. Limitations

There are also important limitations. First, not every user writes frequently. With sparse timelines, temporal signals become fragmented, and the model may make errors or raise warnings too late.

Second, temporal data are often incomplete: users delete posts, switch platforms, and vary their posting frequency. This introduces noise and can hinder learning and generalization.

Thirdly, false alarms are possible. Sometimes a person posts emotionally, jokingly, metaphorically, or reacts to news rather than to personal issues. We tried to reduce this by normalizing and adjusting thresholds, but such errors cannot be completely eliminated.

Also, we primarily used text. But sometimes other signals are important-images, audio, people's reactions, and interactions in comments. Without them, the picture may be incomplete.

In addition, the study was conducted under controlled research conditions; therefore, the reported results should not be interpreted as direct evidence of readiness for unsupervised real-world deployment.

5.4. Ethical Considerations

Mental health is a very sensitive topic, so ethics comes first. Such a system does not replace the doctor and should not make a diagnosis. It is needed for monitoring and scientific tasks, and not for final decisions.

It is important that the person remains "in the process". Any automatic warnings must be verified by specialists so that actions are safe and correct.

Confidentiality is also very important. Data can tell a lot about a person's life over time, so they must be strictly anonymized, stored safely, and comply with the rules, for example, GDPR (General Data Protection Regulation) and the requirements of the platforms themselves [20]. It is also necessary to take into account the right of a person to refuse participation and remember that conclusions about the condition are made even based on open data, so you need to be especially careful with this.

Particular caution is required when analyzing publicly available data, because public accessibility does not automatically imply ethical acceptability for sensitive inference.

Conclusion

In this paper, we presented a temporal deep learning approach that models how depressive and suicidal signals unfold over time in social media. By combining BERT-based embeddings with temporal features and psycholinguistic metrics, the framework goes beyond single-post classification and instead analyzes user-level emotional trajectories. Our results suggest that deterioration-related patterns may become detectable earlier than overt high-risk language, in some cases, days or weeks in advance.

These findings indicate the potential utility of temporal models for monitoring-oriented decision support in digital mental health contexts, provided that any use remains subject to human oversight and careful validation. In large communities, where manual monitoring at scale is unrealistic, real-time risk signals can help prioritize attention and guide follow-up by specialists.

The main result is that the model can detect weak signals that accumulate over time and identify moments when a steady deterioration begins. This is useful for early warning and for support systems that need to respond proactively.

Future work should extend the framework in several directions: incorporating multimodal signals (images, audio, interactions), adapting models to different languages and cultures, and adding social-context modeling via graph neural networks (GNNs) to better capture how environment and communication shape emotional states.

Further validation on diverse, real-world, multilingual datasets is necessary before broader operational use can be considered.

References

[1] Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review: The influence of social media on depression, anxiety, and psychological distress in adolescents. International Journal of Adolescence and Youth, 25(1), 79–93. https://doi.org/10.1080/02673843.2019.1590851

[2] Cheng, J. C., & Chen, A. L. P. (2022). Multimodal time-aware attention networks for depression detection. Journal of Intelligent Information Systems, 59, 319–339. https://doi.org/10.1007/s10844-022-00704-w

[3] Ilias, L., & Askounis, D. (2023). Multitask learning for recognizing stress and depression in social media. Online Social Networks and Media, 37–38, 100270. https://doi.org/10.1016/j.osnem.2023.100270

[4] Chandrasekaran, R., Kotaki, S., & Nagaraja, A. H. (2024). Detecting and tracking depression through temporal topic modeling of tweets: insights from a 180-day study. npj Mental Health Research, 3, 62. https://doi.org/10.1038/s44184-024-00107-5

[5] Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., Kane, J. M., et al. (2020). Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook. Schizophrenia, 6, Article 38. https://doi.org/10.1038/s41537-020-00125-0

[6] Li, Z., An, Z., Cheng, W., Zhou, J., Zheng, F., & Hu, B. (2023). MHA: a multimodal hierarchical attention model for depression detection in social media. Health Information Science and Systems, 11(1), 6. https://doi.org/10.1007/s13755-022-00197-5

[7] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. npj Digital Medicine, 3, 43. https://doi.org/10.1038/s41746-020-0233-7

[8] Seabrook, E. M., Kern, M. L., Fulcher, B. D., & Rickard, N. S. (2018). Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. Journal of Medical Internet Research, 20(5), e9267. https://doi.org/10.2196/jmir.9267

[9] Kerasiotis, M., et al. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining, 14, Article 116. https://doi.org/10.1007/s13278-024-01360-4

[10] Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. (2024). Deep temporal modelling of clinical

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

depression through social media text. Natural Language Processing Journal, 100052. https://doi.org/10.1016/j.nlp.2023.100052

[11] Narynov, S., Mukhtarkhanuly, D., Omarov, B., Kozhakhmet, K., & Omarov, B. (2020). Machine learning approach to identifying depression related posts on social media. In 2020 20th International Conference on Control, Automation and Systems (ICCAS) (pp. 6–11). IEEE. https://doi.org/10.23919/ICCAS50221.2020.9268336

[12] Omarov, B., Narynov, S., & Zhumanov, Z. (2023). Artificial intelligence-enabled chatbots in mental health: A systematic review. Computers, Materials & Continua, 74(3), 5105–5122. https://doi.org/10.32604/cmc.2023.034655

[13] Bokolo, B. G., & Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. Electronics, 12(21), 4396. https://doi.org/10.3390/electronics12214396

[14] Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: Literature review. Interactive Journal of Medical Research, 13, e55067. https://doi.org/10.2196/55067

[15] Malgaroli, M., Hull, T. D., Zech, J. M., et al. (2023). Natural language processing for mental health interventions: A systematic review and research framework. Translational Psychiatry, 13, 309. https://doi.org/10.1038/s41398-023-02592-2

[16] Jaidka, K. (2022). Cross-platform- and subgroup-differences in the well-being effects of Twitter, Instagram, and Facebook in the United States. *Scientific Reports*, 12, 3271. https://doi.org/10.1038/s41598-022-07219-y

[17] Lasri, S., Nfaoui, E. H., & Mrizik, K. (2024). Suicide ideation and risk detection from social media using GPT models. Journal of Computer Science, 20(10), 1349–1356. https://doi.org/10.3844/jcssp.2024.1349.1356

[18] Malhotra, A., & Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. Applied Soft Computing, 130, 109713. https://doi.org/10.1016/j.asoc.2022.109713

[19] Liu, D., Zhang, R., Choo, H., & Li, D. (2022). Detecting and measuring depression on social media using a machine learning approach: Systematic review. JMIR Mental Health, 9(3), e27244. https://doi.org/10.2196/27244

[20] Figuerêdo, J. S. L., Maia, A. L. L. M., & Calumby, R. T. (2022). Early depression detection in social media based on deep learning and underlying emotions. Online Social Networks and Media, 31, 100225. https://doi.org/10.1016/j.osnem.2022.100225