TiC-MuFormer: Time-Aware Caption-Integrated Multimodal Transformers for User-Level Mental Health Modeling

Georgios Tsoumplekas[1], **Yannis Spyridis**[2], **Vasileios Argyriou**[1]

[1]Department of Networks and Digital Media, Kingston University London, UK
[2]Department of Computer Science, Kingston University London, UK
{g.tsoumplekas, y.spyridis, vasileios.argyriou}@kingston.ac.uk

Abstract

User-level affective modeling from social media requires integrating heterogeneous signals that unfold over time. While prior work has focused predominantly on textual analysis, visually expressed affect and temporal posting patterns also carry important psychological cues. However, these modalities are difficult to combine in practice due to sparse emotional evidence, asynchronous posting behavior, and frequent semantic misalignment between images and accompanying text. This paper introduces TiC-MuFormer, a time-enriched caption-integrated multimodal transformer that addresses these challenges by verbalizing visual content through image captioning before fusion and injecting temporal structure prior to cross-modal attention, enabling user trajectories to be modeled in a time-aware semantic space. We instantiate the method on a mental health detection task and demonstrate that it achieves state-of-the-art results across all user-level metrics, outperforming both unimodal and multimodal baselines. Ablation studies further show that temporal coverage, batch size and encoder choice jointly influence downstream accuracy, underscoring the importance of aligned temporal and semantic representations. Overall, this work highlights caption-guided temporal multimodality as a principled modeling strategy for general affective or psychiatric risk inference in social platforms.

Keywords: Multimodal Transformers, Mental Health Detection, Social Media Analysis, Image Captioning, Temporal Modelling

1. Introduction

Mental wellbeing increasingly manifests through online behaviour, as users regularly disclose affective states, coping strategies, or emotional fluctuations on social media platforms (De Choudhury et al., 2013; Coppersmith et al., 2014). Because such platforms provide large-scale, longitudinal traces of self-expression, they offer a unique lens into behavioural and psychological patterns that may not be observable in traditional clinical settings. As a result, modelling affective signals from user-generated content has emerged as a key research area in computational social science and digital mental health analytics.

While early work focused predominantly on textual content, multimodal affect modelling is now recognized as essential, as many psychologically relevant cues appear visually rather than linguistically (Reece and Danforth, 2017). For example, colour palette, composition, or symbolism in posted images may convey emotional nuance that is absent from the accompanying text. Beyond multimodality, temporal structure plays a crucial role since user-level signals often arise not from isolated posts but from how linguistic and visual expression evolves over time (Bucur et al., 2023; Cheng and Chen, 2022).

However, jointly modelling these dimensions poses challenges. First, affective cues are sparse and unevenly distributed, with long stretches of neutral content overshadowing sporadic psychologically-relevant posts. Second, posting behaviour is asynchronous, making chronological distance between posts itself an informative variable. Third, text–image pairs are frequently semantically misaligned, as short conversational text fragments may under-specify the emotional content depicted in an image. Conventional multimodal transformers fuse the two modalities but do not repair this semantic gap. Additionally, they do not explicitly integrate temporal dynamics before cross-modal interaction.

In this paper, we introduce `TiC-MuFormer`, a time-enriched caption-integrated multimodal transformer that addresses these limitations through two synergistic mechanisms: (i) image captioning is used to verbalize visual cues before fusion, aligning the two modalities in a linguistically coherent space and (ii) temporal embeddings are injected prior to cross-modal attention, allowing behavioural regularities to shape semantic alignment. Although our experiments instantiate the approach on depression detection, the architecture is designed to be task-agnostic and could in principle support broader user-level affective or wellbeing inference tasks. However, the present study evaluates it only on the Twitter mental health dataset (Shen et al.,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9669–9677
11–16 May 2026.

2017), where TiC-MuFormer achieves state-of-the-art performance across all metrics, underscoring the importance of both caption-guided multimodal alignment and temporal integration. The main contributions of this work are as follows:

- We introduce TiC-MuFormer, a time-enriched caption-integrated multimodal transformer that unifies textual, visual, and temporal information at the user level, and whose design is task-agnostic and potentially applicable beyond depression detection.
- We achieve new state-of-the-art results on the Twitter mental health detection benchmark, surpassing all existing unimodal and multimodal baselines across standard evaluation metrics.
- We provide a detailed ablation study demonstrating how temporal coverage, batch size, and encoder choice systematically contribute to performance and offer empirical insight into why the architecture is effective.

2. Related Work

2.1. Text-based Affective State Modelling

Early approaches to modelling affective states in online text primarily relied on feature engineering and shallow learning architectures. Typical pipelines combined bag-of-words or n-gram representations with linear classifiers, while psycholinguistic lexicons such as LIWC were employed to characterize affective tone and psychological markers of distress (Pennebaker et al., 2015). Topic-model–based approaches, most notably LDA, also played a foundational role in uncovering latent affective themes (Blei et al., 2003). Collectively, these methods demonstrated that linguistic regularities in social media discourse correlate with self-reported mental states and facilitated the first large-scale wellbeing monitoring efforts using user-generated content (De Choudhury et al., 2013).

With the emergence of deep learning, CNN- and RNN-based encoders improved sensitivity to subtle affective cues by learning distributed semantic representations directly from text (Kim, 2014), although their performance remained constrained by fixed context windows. The introduction of transformer architectures and contextualized language models marked a substantial shift since pre-trained encoders such as RoBERTa enabled richer modelling of affective nuance by capturing long-range dependencies and conversational structure (Liu et al., 2019). Emotion-aware variants, such as EmoBERTa, advanced this trajectory further by explicitly incorporating affective and conversational signals into the embedding space (Kim and Vossen, 2021). More recently, large language models (LLMs) have exhibited meaningful zero- and few-shot capabilities for wellbeing-related inference from text, such as clinical-style interview settings, reinforcing the view that contextual, emotion-informed representations are central to affective modelling (Ohse et al., 2024).

2.2. Multimodal Modelling of Affective Signals

Text-only approaches overlook affective cues expressed visually. Prior studies show that image properties such as colourfulness, saturation, and facial composition correlate with emotional state (Reece and Danforth, 2017), motivating multimodal user-level modelling. The first large-scale benchmark incorporating both text and images was introduced by Shen et al. (2017), who integrated linguistic, behavioural, and visual features using multimodal dictionary learning. Subsequent work explored reinforcement-learning–based post selection to summaries long user timelines (Gui et al., 2019b), improving robustness by prioritizing psychologically salient signals.

More recent work improves the quality of visual representations by adopting self-supervised and contrastive encoders such as CLIP (Radford et al., 2021), which transfer more effectively to user-generated content than supervised backbones (Ericsson et al., 2021). In parallel, multimodal captioning methods (Srivatsan et al., 2024; Patel et al., 2025) enhance semantic alignment across modalities by grounding image content in natural language, enabling finer-grained fusion than simple embedding concatenation. Collectively, these advances have shifted multimodal affect modelling toward architectures that integrate vision and language through jointly aligned representations rather than through loosely coupled parallel streams.

2.3. Temporal and Behavioral Dynamics in User-Level Mental State Inference

Earlier work has explored cooperative multimodal agents that first select psychologically salient posts prior to fusion (Gui et al., 2019b) as well as topic-enriched auxiliary learning frameworks that inject modality-agnostic topic cues to stabilize long-horizon user inference (An et al., 2020). More recent research demonstrates that user-level mental state inference on social platforms further benefits from explicit temporal modeluing and behaviour-aware multimodal fusion. Time-enriched transformers, such as Time2Vec (Bucur et al., 2023), encode irregular posting intervals and integrate text–image signals at the user level, yielding substantial gains over post-level baselines. Complementary architectures based on time-aware atten-

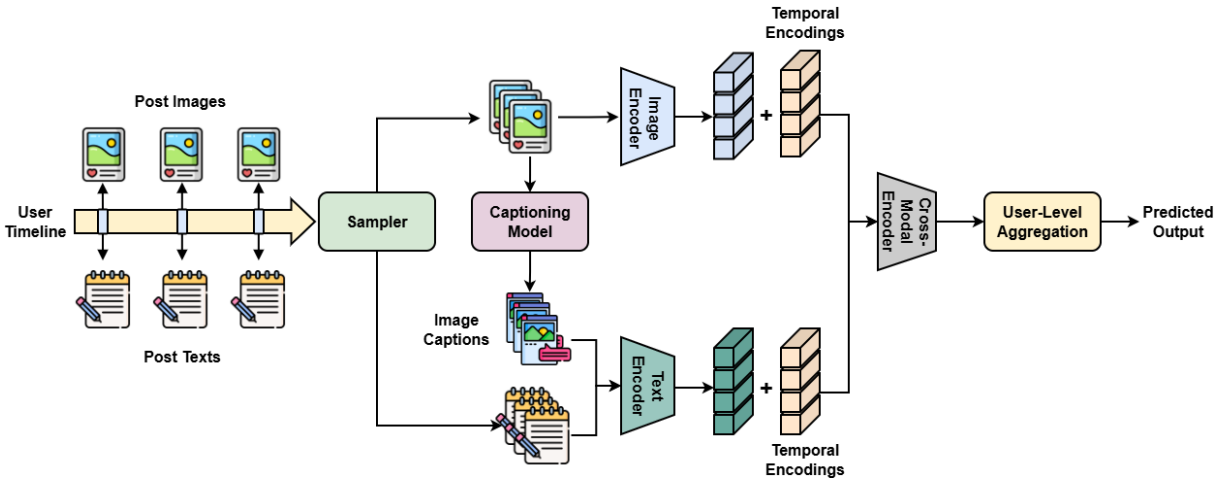

Figure 1: Overview of the `TiC-MuFormer` architecture.

tion similarly combine text, images, and posting cadence through time-aware LSTMs augmented with post-importance attention (Cheng and Chen, 2022), underscoring that temporal structure carries affective signal beyond the content itself.

On the text side, hybrid CNN–BiLSTM architectures with attention mechanisms capture both local lexical patterns and long-range dependencies, enabling competitive user-level screening in the absence of visual information (Thekkekara et al., 2024). On the vision–language side, attention-based fusion models integrate implicit visual semantics with explicit textual content, improving robustness in scenarios where user text is sparse or off-topic (Long et al., 2023).

Beyond content, personality- and sentiment-aware ensembles introduce interpretable behavioral priors into the fusion process (Pradnyana et al., 2025a), while context-driven multimodal networks incorporate learned context vectors to adaptively reweigh modalities over time (Tahir et al., 2025). Finally, domain-specific pretraining, as in MentalBERT and MentalRoBERTa, equips language encoders with sensitivity to mental-health discourse, yielding consistent improvements in downstream user-level screening (Ji et al., 2022).

3. Methodology

3.1. Problem Formulation and Model Overview

We consider the task of user-level emotion recognition from a sequence of multimodal social media posts. Let a user be represented by an asynchronous timeline:

$$\mathcal{U} = \left\{ \left(\mathbf{m}_k^{(\text{text})},\ \mathbf{m}_k^{(\text{img})},\ t_k \right) \right\}_{k=1}^{N_{\mathcal{U}}} \tag{1}$$

where $\mathbf{m}_k^{(\text{text})}$ denotes the textual component of the k-th post, $\mathbf{m}_k^{(\text{img})}$ the associated image, and t_k the corresponding posting timestamp. The objective is to predict a user-level binary label $y_{\mathcal{U}} \in \{0, 1\}$ indicating the presence or absence of an emotion of interest. The model therefore implements a mapping $f_\Theta : \mathcal{U} \to [0, 1]$ where Θ denotes all trainable parameters, and optimization is performed using binary cross-entropy.

A user timeline may contain thousands of posts, most of which are not predictive for the final user-level classification. To handle this efficiently, we operate on fixed-length subsequences defined as:

$$\mathcal{S}_{\mathcal{U}} = \left\{ \left(\mathbf{m}_k^{(\text{text})},\ \mathbf{m}_k^{(\text{img})},\ t_k \right) \mid k \in [s, s + K - 1] \right\} \tag{2}$$

where K is the subsequence window size and s is a sampled start index. Each $\mathcal{S}_{\mathcal{U}}$ therefore corresponds to a temporally local slice of user behaviour. During training, a mini-batch is formed as $\mathcal{B} = \{\mathcal{S}_{\mathcal{U}_1}, \ldots, \mathcal{S}_{\mathcal{U}_B}\}$ with one subsequence per user. A user-level embedding is later derived by aggregating information across the posts in $\mathcal{S}_{\mathcal{U}}$ after multimodal fusion and temporal encoding.

Our proposed architecture, `TiC-MuFormer` (Time-and-Caption Multimodal Transformer), builds directly on this formulation by combining caption-guided enrichment of textual representations, which preserves visual semantics during fusion and temporal encoding that captures behavioural regularities and irregularities in posting activity. These components jointly enable the model to represent both what the user expresses and how those expressions evolve over time which is essential for reliable user-level emotion inference. Figure 1 illustrates this overall workflow, showing how user timelines are transformed into temporally enriched multimodal representations prior to user-level classification.

3.2. Caption-Augmented Multimodal Representation Learning

A core challenge is cross-modal asymmetry, as emotionally relevant cues are often conveyed through the image while the accompanying text is short or generic. Without additional mediation, these signals remain inaccessible to the language stream. To bridge this gap, we convert each image into a natural-language description using an instruction-tuned captioning model $\mathcal{G}_\phi$, ensuring that salient visual content is verbalized and can inform downstream reasoning even when the original text is sparse or semantically detached. For a post indexed by k, we define:

$$c_k = \mathcal{G}_\phi\big(\mathbf{m}_k^{(\text{img})}\big), \tag{3}$$

where c_k is the generated caption and $\mathbf{m}_k^{(\text{img})}$ denotes the raw image. The caption is then concatenated with the original text to form an enriched textual input:

$$\tilde{\mathbf{m}}_k^{(\text{text})} = \text{concat}\left(\mathbf{m}_k^{(\text{text})},\ c_k\right), \tag{4}$$

ensuring that visual context becomes explicitly available to the language encoder.

The enriched text is encoded using a pre-trained language model $\mathcal{E}_{\text{text}}$, and the image is embedded using a pre-trained vision encoder $\mathcal{E}_{\text{img}}$:

$$\begin{aligned} \mathbf{h}_k^{(\text{text})} &= \mathcal{E}_{\text{text}}\big(\tilde{\mathbf{m}}_k^{(\text{text})}\big) \in \mathbb{R}^{d_t}, \\ \mathbf{h}_k^{(\text{img})} &= \mathcal{E}_{\text{img}}\big(\mathbf{m}_k^{(\text{img})}\big) \in \mathbb{R}^{d_v} \end{aligned} \tag{5}$$

The two embeddings are projected into a shared latent space of dimension d via learned linear mappings:

$$\mathbf{z}_k^{(\text{text})} = \mathbf{W}_t \mathbf{h}_k^{(\text{text})}, \qquad \mathbf{z}_k^{(\text{img})} = \mathbf{W}_i \mathbf{h}_k^{(\text{img})} \tag{6}$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times d_t}$ and $\mathbf{W}_i \in \mathbb{R}^{d \times d_v}$. When no image is present for a given post, the corresponding visual token is masked so that it does not contribute to attention-based fusion downstream.

3.3. Temporal Encoding and User-Level Inference

Since user posting behavior is asynchronous, temporal structure is incorporated through a Time2Vec (Kazemi et al., 2019) embedding. Each timestamp t_k is mapped to a temporal vector $\mathbf{p}_k \in \mathbb{R}^{d_\tau}$ via:

$$\begin{aligned} \text{T2V}(t_k)[r] &= \begin{cases} \omega_r\, g(t_k) + \varphi_r, & r = 0 \\ \sin\big(\omega_r\, g(t_k) + \varphi_r\big), & 1 \le r \le d_\tau \end{cases} \\ g(t_k) &= \frac{1}{t_k + \epsilon} \end{aligned} \tag{7}$$

where ω_r and φ_r are learnable parameters, d_τ is the temporal embedding dimension, and $\epsilon > 0$ prevents division by zero.

This temporal signal is injected into both modality streams to produce time-aware embeddings:

$$\tilde{\mathbf{z}}_k^{(\text{text})} = \mathbf{z}_k^{(\text{text})} + \mathbf{p}_k, \qquad \tilde{\mathbf{z}}_k^{(\text{img})} = \mathbf{z}_k^{(\text{img})} + \mathbf{p}_k \tag{8}$$

A cross-modal transformer $\mathcal{F}_{\text{cross}}$ then fuses the representations:

$$\mathbf{z}_k = \mathcal{F}_{\text{cross}}\big(\tilde{\mathbf{z}}_k^{(\text{text})},\ \tilde{\mathbf{z}}_k^{(\text{img})}\big) \tag{9}$$

The sequence $\{\mathbf{z}_k\}$ is subsequently processed by a temporal transformer with multi-head self-attention, and a user-level embedding is obtained via mean pooling. Finally, a logistic classifier produces the user-level prediction.

4. Experimental Results

4.1. Experimental Setting

Dataset. Experiments are conducted on the widely used multimodal Twitter depression dataset introduced by Shen et al. (2017). The dataset consists of 1402 users labelled as self-reporting a diagnosis of depression and 1402 control users, with all posts collected over a one-month period surrounding a clinically indicative anchor tweet. Following prior work, we adopt the evaluation protocol of Bucur et al. (2023), applying five-fold cross-validation with identical train/test splits to ensure comparability across methods.

Model Configuration. The proposed architecture employs Llama 3.2-Vision (Grattafiori et al., 2024) for caption generation, RoBERTa (Liu et al., 2019) as the language encoder, CLIP (Radford et al., 2021) for visual embeddings, and an LXMERT-style cross-modal transformer (Tan and Bansal, 2019) for early fusion. Sequential posts are processed using subsequence sampling with a window size of $K = 512$ and a batch size of 128. All models are trained using the Adam optimizer with a base learning rate of 10^{-5}, adjusted via a cyclical scheduler that varies between 10^{-5} and 10^{-4} over 10 epochs. The fusion module comprises four cross-encoder layers with eight attention heads and embedding dimension $d = 128$, while the temporal transformer consists of two layers with identical configuration. During inference, 10 subsequences are sampled per user, and the final prediction is obtained through majority voting.

Baselines. We compare against two groups of baselines models. The text-only group includes T-LSTM (Baytas et al., 2017), EmoBERTa (Kim and Vossen, 2021), LSTM+RL (Gui et al., 2019a), and CNN+RL (Gui et al., 2019a). The multimodal group includes MTAL (An et al., 2020) ,

Table 1: Comparison of unimodal (text) and multimodal (text + image) models on the Twitter Depression Dataset (Gui et al., 2019b) (T: Text, I: Images, C: Captions).

Model	Modality	F1	Precision	Recall	Accuracy
T-LSTM (Baytas et al., 2017)	T	0.848	**0.896**	0.804	0.855
EmoBERTa Transformer	T	0.864	0.843	0.887	0.861
LSTM + RL (Gui et al., 2019a)	T	0.871	0.872	0.870	0.870
CNN + RL (Gui et al., 2019a)	T	0.871	0.871	0.871	0.871
TiC-MuFormer (Ours)	T+C	**0.880**	0.864	**0.896**	**0.878**
MTAL (An et al., 2020)	T+I	0.842	0.842	0.842	0.842
GRU + VGG-Net + COMMA (Gui et al., 2019b)	T+I	0.900	0.900	0.901	0.900
MTAN (Cheng and Chen, 2022)	T+I	0.908	0.885	0.931	–
Vanilla Transformer (Bucur et al., 2023)	T+I	0.886	0.868	0.905	0.883
Set Transformer (Bucur et al., 2023)	T+I	0.902	0.878	0.928	0.924
Time2Vec Transformer (Bucur et al., 2023)	T+I	0.931	0.931	0.931	0.931
MTEN (Zafar et al., 2024)	T+I	0.945	0.945	0.945	0.945
Chat-Diagnose (Qin et al., 2025)	T+I	0.946	**0.979**	0.915	0.948
DeXMAG (Pradnyana et al., 2025b)	T+I	0.952	0.961	0.957	0.959
ContextVecNet (Tahir et al., 2025)	T+I	0.962	0.977	0.948	0.963
TiC-MuFormer (Ours)	**T+I+C**	**0.971**	0.969	**0.972**	**0.971**

GRU+VGGNet+COMMA (Gui et al., 2019b), MTAN (Cheng and Chen, 2022), Vanilla/Set/Time2Vec Transformers (Bucur et al., 2023), MTEN (Zafar et al., 2024), Chat-Diagnose (Qin et al., 2025), DeXMAG (Pradnyana et al., 2025a), and ContextVecNet (Tahir et al., 2025), covering both established and recent approaches in multimodal user-level inference for depression detection.

Recent multimodal large language models, such as LLaMA-Vision, can in principle be fine-tuned for multimodal reasoning tasks. However, our objective differs from general multimodal generation or question answering: the task requires modelling long user timelines with potentially thousands of posts while capturing irregular temporal dynamics. Direct fine-tuning of large multimodal LLMs on such long behavioural sequences is computationally demanding and does not naturally incorporate explicit representations of posting intervals or temporal trajectories. TiC-MuFormer instead separates representation learning into caption-guided semantic alignment and explicit temporal encoding through Time2Vec, enabling efficient modelling of long asynchronous user timelines. In this sense, the proposed approach should be viewed as a timeline-oriented multimodal modelling framework rather than a replacement for general multimodal LLMs.

Evaluation Metrics. Performance is reported using Accuracy, Precision, Recall, and F1 score, consistent with prior work on this benchmark.

4.2. Main Results

Table 1 presents the performance of `TiC-MuFormer` alongside unimodal and multimodal baselines. In the unimodal (text-only) setting, `TiC-MuFormer` achieves an F1 score of 0.880, Recall of 0.896, and Accuracy of 0.878, all of which surpass those of competing text-only models. Although T-LSTM attains slightly higher Precision (0.896 vs. 0.864), it does so at the cost of Recall (0.804), leading to a lower overall F1 score (0.848). This demonstrates that our caption-enriched text representation establishes a more balanced and robust decision boundary, enhancing the model's ability to identify positive cases which was the key weakness of earlier unimodal approaches.

In the full multimodal setting, `TiC-MuFormer` attains an F1 score of 0.971, outperforming the next best model (ContextVecNet, 0.962) by 0.9 percentage points. Recall likewise increases from 0.948 (ContextVecNet) to 0.972, and Accuracy from 0.963 to 0.971. Across all metrics except Precision, our method is the top-performing model. While `TiC-MuFormer` (0.969) trails Chat-Diagnose (0.979) and ContextVecNet (0.972) by a narrow margin in Precision, it still exceeds all other prior models and achieves the highest Recall overall. It is worth noting that, for early-screening tasks, where missing positive cases is especially costly, this recall advantage is more critical than marginal gains in precision.

From the obtained results, the performance gains appear to be associated with the interaction of caption-guided semantic enrichment and temporally informed fusion. Captioning likely helps bridge cross-modal asymmetry by converting implicit visual cues into explicit linguistic structure, while Time2Vec enables the model to exploit temporal regularities and behavioural patterns that token-level text encoders overlook. Together, these com-

Table 2: Ablation summary in the multimodal setting. Each row varies one component while holding the others fixed. We report the F1 score obtained in each setting.

Component Varied	Setting	F1	Best Config
Window size K	32	0.887	RoBERTa, B=128
	64	0.922	RoBERTa, B=128
	128	0.936	RoBERTa, B=128
	256	0.960	RoBERTa, B=128
	512	**0.971**	RoBERTa, B=128
Batch size B	**128**	**0.971**	RoBERTa, K=512
	256	0.964	RoBERTa, K=512
Text encoder	**RoBERTa**	**0.971**	K=512, B=128
	EmoBERTa	0.956	K=512, B=128

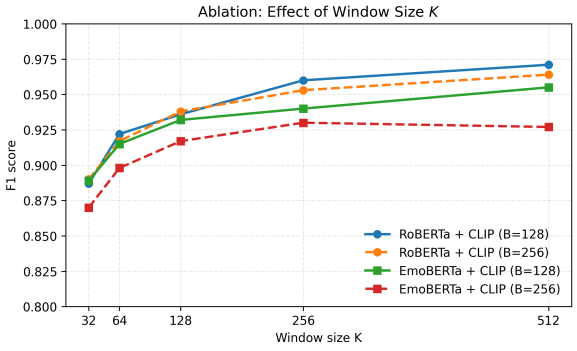

Figure 2: Effect of temporal window size K on F1 performance for different text encoders and batch sizes.

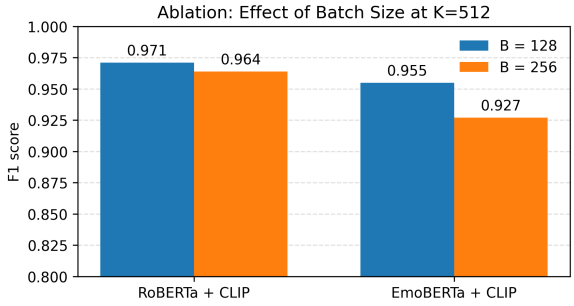

Figure 3: Comparison of batch sizes at the optimal window length $K = 512$ for RoBERTa and EmoBERTa encoders.

ponents yield consistent and statistically meaningful improvements, establishing `TiC-MuFormer` as the new state of the art on this benchmark.

4.3. Ablation Studies

We conduct a controlled ablation to quantify the effect of (i) the temporal window size K, (ii) the batch size B, and (iii) the underlying text encoder. All experiments are carried out in the multimodal setting using CLIP for image embeddings and Time2Vec for temporal encoding.

Effect of Temporal Window Size. Table 2 and Figure 2 show that performance improves consistently as K increases. The F1 score rises from 0.887 at K=32, to 0.922 at K=64, 0.936 at K=128, 0.960 at K=256, and peaks at 0.971 for K=512 (using RoBERTa with B=128). This corresponds to an 8.4% gain relative to the smallest window size, indicating that long-range temporal context substantially enhances user-level modeling.

Effect of Batch Size. The influence of batch size is summarized in Table 2 and visualized in Figure 3. The best configuration uses B=128 (F1=0.971), while increasing the batch size to 256 leads to a small but systematic decrease (0.964 for RoBERTa and 0.927 for EmoBERTa). The line plot in Figure 2 further confirms this trend since the solid traces (B=128) dominate the dashed ones (B=256) across all window sizes suggesting that smaller batches preserve higher temporal sensitivity, while larger batches oversmooth gradients and reduce responsiveness to posting dynamics.

Effect of Text Encoder. Finally, we evaluate the impact of the textual encoder. As reported in Table 2, RoBERTa achieves the strongest overall performance (F1=0.971 at K=512, B=128), surpassing EmoBERTa (0.955) under the same hyperparameters. Since caption-guided enrichment already surfaces emotional signals linguistically, the model benefits more from RoBERTa's general-purpose semantic expressiveness than from additional emotion-specialized pretraining.

Figure 2 also shows that RoBERTa consistently outperforms EmoBERTa across all values of K. This is likely because captioning already embeds affective cues into the textual stream, reducing the marginal benefit of emotion-specific pretraining and favoring RoBERTa's broader semantic coverage.

5. Conclusion

This paper introduced `TiC-MuFormer`, a time-enriched caption-integrated multimodal transformer for user-level affective modelling from social media. The architecture aligns visual and textual content via captioning and incorporates temporal structure prior to fusion, enabling the model to jointly encode what users express and how these signals evolve over time. Although our experiments instantiate the framework on mental health detection, the method is designed to be task-agnostic and could potentially generalize to other forms of behavioural or psychiatric risk profiling. However, broader validation across datasets remains future work.

Empirical evaluation on the Twitter Dataset demonstrates that `TiC-MuFormer` achieves state-of-the-art performance, outperforming prior unimodal and multimodal baselines across user-level metrics. The ablation study further shows that both semantic realignment through captioning and temporal enrichment via Time2Vec are key contributors to this improvement. The broader implication of this work is that multimodal temporal fusion is a principled modelling strategy for digital mental health analytics extending beyond mental health to related affective and wellbeing signals such as anxiety, loneliness, relapse risk, or emotional instability.

6. Ethics Statement and Limitations

The model is trained on a dataset whose user population is demographically narrow, with a strong over-representation of young US-based users. As a result, the learned decision boundaries may not transfer uniformly to other demographic or cultural groups, reflecting well-documented risks of exclusion and representational bias in social media–sourced datasets Hovy et al. (2016). Furthermore, the supervision signal relies on self-reported labels rather than clinically verified diagnoses meaning that the timing and persistence of the referenced mental health condition are unknown. Prior work has shown that linguistic and behavioural markers can shift following treatment or recovery (Harrigian and Dredze, 2022), which limits causal interpretation of model outputs. Future work should therefore evaluate the proposed architecture across additional datasets, platforms, and mental health conditions to better assess its robustness and generalisability.

Another limitation concerns the captioning stage used to verbalise image content. While captioning helps bridge the semantic gap between visual and textual modalities, the generated descriptions may reflect biases or priors from the underlying captioning model and may occasionally introduce hallucinated or emotionally loaded language. As a result, part of the observed performance gains may stem from information injected by the captioning model rather than solely from improved multimodal alignment. A deeper analysis of caption quality, bias propagation, and hallucination effects remains an important direction for future work.

It is also important to emphasise that this system is not designed to provide a clinical assessment. The predictions capture possible behavioural indicators, not diagnostic outcomes, and any medical interpretation must remain the responsibility of qualified health professionals. Consequently, deployment in real-world settings should be accompanied by caution regarding fairness, interpretability, and appropriate downstream use.

Acknowledgements

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135800 (RAIDO).

References

Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics*, pages 1078–1089.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74.

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

Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. 2023. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European conference on information retrieval*, pages 200–215. Springer.

Ju Chun Cheng and Arbee LP Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, 59(2):319–339.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology:*

From linguistic signal to clinical reality, pages 51–60.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. 2021. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5414–5423.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. 2019a. Depression detection on social media with reinforcement learning. In *China National Conference on Chinese Computational Linguistics*, pages 613–624. Springer.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019b. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 110–117.

Keith Harrigian and Mark Dredze. 2022. Then and now: Quantifying the longitudinal validity of self-disclosed depression diagnoses. *arXiv preprint arXiv:2206.11155*.

Dirk Hovy, Shannon L Spruit, et al. 2016. The social impact of natural language processing. In *The 54th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference, Vol. 2 (Short Papers)*. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. 2019. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xinping Long, Yifan Zhang, Xin Shu, and Jian Shu. 2023. Image-text fusion model for depression tendency detection based on attention. In *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 730–734. IEEE.

Julia Ohse, Bakir Hadžić, Parvez Mohammed, Nicolina Peperkorn, Michael Danner, Akihiro Yorita, Naoyuki Kubota, Matthias Rätsch, and Youssef Shiban. 2024. Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech & Language*, 88:101663.

Vrajkumar Patel, Aayush Modi, Harsh Mistry, Abhishesh Mishra, Rocky Upadhyay, and Apoorva Shah. 2025. From alt-text to real context: Revolutionizing image captioning using the potential of llm. *IJS-CSEIT*, 11(1):379–387.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Gede Aditra Pradnyana, Wiwik Anggraeni, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo. 2025a. An explainable ensemble model for revealing the level of depression in social media by considering personality traits and sentiment polarity pattern. *Online Social Networks and Media*, 46:100307.

Gede Aditra Pradnyana, Wiwik Anggraeni, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo. 2025b. Revealing depression through social media via adaptive gated cross-modal fusion augmented with insights from personality traits. *IEEE Access*.

Wei Qin, Zetong Chen, Xun Yang, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2025. Explainable and interactive llms-augmented depression detection in social media. *IEEE Transactions on Computational Social Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, volume 2017, pages 3838–3844.

Nikita Srivatsan, Sofia Samaniego, Omar Florez, and Taylor Berg-Kirkpatrick. 2024. Alt-text with context: Improving accessibility for images on twitter. In *The Twelfth International Conference on Learning Representations*.

Waleed Bin Tahir, Shah Khalid, Saied Alshahrani, Shuaa S Alharbi, and Haifa F Alhasson. 2025. Contextvecnet: A context-driven multimodal learning framework for depression detection. *IEEE Journal of Biomedical and Health Informatics*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Joel Philip Thekkekara, Sira Yongchareon, and Veronica Liesaputra. 2024. An attention-based cnn-bilstm model for depression detection on social media text. *Expert systems with applications*, 249:123834.

Anas Zafar, Danyal Aftab, Rizwan Qureshi, Yaofeng Wang, and Hong Yan. 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*, pages 2258–2265.