

Article

DR-Transformer: A Dual-Regularized Transformer Combining Sparse Attention and Supervised Contrastive Learning for Interpretable Stress Detection in Social Media Text

Mehdi Chrifi Alaoui , Nour-Eddine Joudar and Mohamed Ettaouil

Faculty of Sciences and Techniques, University Sidi Mohammed Ben Abdellah, Fez 30000, Morocco; nour-eddine.joudar@usmba.ac.ma (N.-E.J.); mohamed.ettaouil@usmba.ac.ma (M.E.)

* Correspondence: mehdi.chrifialaoui@usmba.ac.ma

Abstract

Automatic detection of stress in social media text holds promise for supporting digital mental health, but most existing Transformer-based approaches are opaque and computationally demanding. This work presents DR-Transformer, a Dual-Regularized Transformer that combines two complementary mechanisms: (i) a group sparsity penalty ($L_{2,1}/L_2$ elastic net) applied to the query and key projection matrices of every attention head, which encourages whole-row sparsity, producing more concentrated and inspectable attention patterns; (ii) a supervised contrastive loss on the [CLS] projection, which organizes the latent space according to the stress label. The architecture is intentionally lightweight (six layers, eight heads, 256-dim embeddings; ~ 9.5 M parameters) and runs entirely on consumer-grade hardware (NVIDIA GTX 1660, 6 GB). Experiments on the publicly available Dreddit dataset (binary stress classification, 2838 train/715 test segments) compare DR-Transformer against Logistic Regression, BiLSTM, a Standard Transformer of identical architecture, and MentalBERT. Across five seeded runs, DR-Transformer (Full) reaches $F_1 = 0.876$ (bootstrap 95% CI 0.852–0.898), outperforming the Standard Transformer ($F_1 = 0.842$; McNemar $p < 0.001$ with Bonferroni correction) and performing comparably to the much larger MentalBERT ($F_1 = 0.879$; $p = 0.421$). Sparse regularization increases the fraction of near-zero attention weights (below 0.01) from 0.215 to 0.682, while the supervised contrastive loss improves the silhouette score of [CLS] embeddings from 0.312 to 0.483. Dual regularization thus combines accuracy, efficiency, and structurally induced attention concentration in a single model which can be trained without specialized infrastructure. We use the term “interpretable” throughout in this restricted, structural sense—to refer to concentrated and inspectable attention—rather than in the sense of established causal or mechanistic faithfulness; this is only partially and indirectly supported by our token deletion analysis.



Academic Editors: Gianni D’Angelo and Olufemi A. Omitaomu

Received: 17 May 2026

Revised: 28 June 2026

Accepted: 8 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: stress detection; mental health; social media; Transformer; sparse attention; supervised contrastive learning; attention inspectability; computational efficiency; natural language processing; Dreddit

1. Introduction

The proliferation of mental health discourse on platforms such as Reddit, Twitter, and Facebook has created a real-time data source for public health research [1]. Natural language processing (NLP) and deep learning have shown promise in automatically identifying signs of stress, depression, anxiety, and suicidal ideation from such text [2–4]. Stress in particular

is a near-universal experience, the detection of which at scale could support targeted wellbeing interventions, content moderation, and epidemiological monitoring.

Despite this promise, two practical barriers limit potential use in clinical or content-moderation settings. First, the opacity of large Transformer models limits trust and accountability [5,6]; clinicians and content moderators need to understand why a system flags a post before acting on it. Second, many state-of-the-art models—ranging from MentalBERT [7] to MentaLLaMA [8]—are expensive to train and serve, restricting their use in resource-constrained settings.

We propose DR-Transformer, a Dual-Regularized Transformer that targets both barriers. Our contributions are as follows:

1. A **group-sparse attention** regularizer ($L_{2,1}/L_2$ elastic net) on the query and key projection matrices that pushes whole input dimensions toward zero, yielding more concentrated and inspectable attention patterns—a structural property we distinguish from full faithfulness, which would require additional perturbation-based evaluation (Section 4.3).
2. A **supervised contrastive loss** [9] on the [CLS] projection that organizes the latent space by class, improving generalization and visual separability.
3. A **lightweight design** (six layers, eight heads, ~ 9.5 M parameters) that we train and evaluate end-to-end on a single NVIDIA GTX 1660 (6 GB).
4. An empirical study on Dreddit [10] with five seeded runs, bootstrap confidence intervals, McNemar significance tests (reported both per seed and pooled), and ablations isolating each regularizer.

We note that group sparsity penalties and supervised contrastive losses are individually established techniques. The novelty of DR-Transformer lies in three specific aspects: (i) the application of an $L_{2,1}/L_2$ group elastic-net penalty directly to the query and key projection matrices (W_Q, W_K)—rather than to activations or classifier outputs—which induces row-level sparsity over input dimensions and propagates to token-level attention weights. (ii) The joint optimization of both regularizers within a single lightweight encoder trained from scratch without pretrained weights. (iii) The empirical demonstration, with rigorous statistical testing and ablations, that this combination achieves a competitive accuracy–efficiency–inspectability trade-off on the mental health stress detection task, a high-stakes domain where computational accessibility and model transparency are both critical.

Section 2 reviews related work. Section 3 describes the architecture and training. Section 4 reports the classification, interpretability, and efficiency results. Section 5 discusses implications and limitations, and Section 6 concludes.

2. Related Work

2.1. Mental Health Detection from Social Media

Early work used lexicon-based features and classical classifiers [2,3]. CLPsych shared tasks [11,12] and the Dreddit benchmark [10] established standard datasets for distinguishing distressed users from control users. Recent deep learning approaches include time-aware attention for suicidal ideation detection [13], dual-context language models [14], and knowledge-aware attention [15]. Domain-adapted pretrained encoders—most notably MentalBERT [7]—substantially improve downstream performance by continued pretraining on mental health subreddits. The latest wave fine-tunes large language models for mental health analysis, e.g., MentaLLaMA [8], but their multi-billion-parameter scale is incompatible with consumer hardware.

More recently, Kyrou et al. [16] provided a comprehensive survey of deep learning approaches for stress detection, highlighting the ongoing challenge of balancing model capacity with interpretability and deployment feasibility. Soufleri and Ananiadou [17]

proposed a multimodal framework that fuses synthesized visual representations with text for stress detection on Dreddit, achieving strong accuracy but at significantly higher computational costs compared to text-only approaches. Zhuang et al. [18] introduced a BERT-fused model combining LDA topic features with contextual embeddings for psychological stress detection from social media, demonstrating the benefit of integrating structural topic information. Compared to these concurrent approaches, DR-Transformer distinguishes itself by simultaneously targeting attention concentration and computational efficiency within a single lightweight model trainable on consumer hardware—a dimension that remains underexplored in the recent literature.

2.2. Sparse and Efficient Attention

Efficient-attention methods aim to lower the $\mathcal{O}(N^2)$ cost of self-attention. Longformer [19] and BigBird [20] use structured sparsity (windowed plus global tokens). Reformer [21] replaces dot-product attention with locality-sensitive hashing. These approaches target *computational* sparsity. Our work instead induces sparsity on the *projection matrices* (W_Q , W_K) via a group $L_{2,1}$ penalty, targeting *attention concentration for inspection*. The two strategies are orthogonal and could be combined.

2.3. Contrastive Learning for Text

Self-supervised contrastive learning [22–24] produces high-quality sentence embeddings without labels. Supervised contrastive learning [9] extends the framework to labeled data, pulling same-class samples together and pushing different-class apart, and has been shown to outperform pure cross-entropy on text classification [25]. We adopt this objective on the [CLS] projection of our Transformer, providing a complementary signal to the sparse attention regularizer.

3. Materials and Methods

3.1. Overall Architecture

DR-Transformer is an encoder-only Transformer [26] that produces (i) a class logit and (ii) a normalized contrastive embedding from the same [CLS] representation. The pipeline consists of four stages: input text is tokenized with BERT WordPiece [27]; embeddings are added to learned positional encodings; a stack of $L = 6$ Transformer layers processes the sequence, where each layer's attention is regularized by a group elastic-net penalty on W_Q and W_K ; the final [CLS] embedding feeds a binary classifier and a 2-layer MLP projection used by the supervised contrastive loss. Figure 1 provides a schematic overview.

3.2. Input Representation

Let a post be tokenized into a sequence of WordPiece tokens. A special classification token [CLS]—the final hidden state of which serves as the aggregate sequence representation for classification—is prepended, and a separator token [SEP] is appended; sequences are padded or truncated to length $N = 128$ (95th percentile of Dreddit segment lengths). The input is:

$$X = [e_{[\text{CLS}]}, e_1, \dots, e_N, e_{[\text{SEP}]}] + P, \quad (1)$$

where $e_i \in \mathbb{R}^{d_{\text{model}}}$ is a token embedding and P is a learned positional encoding matrix.

Token embeddings are randomly initialized from $\mathcal{N}(0, 0.02)$ and trained from scratch jointly with all other model parameters. DR-Transformer does *not* use pretrained embeddings from BERT or any other external model; the WordPiece vocabulary (30,522 tokens) and tokenizer are reused from BERT for compatibility, but the embedding matrix is not initialized from BERT weights. This distinguishes DR-Transformer from fine-tuning approaches such as MentalBERT. The total parameter count of $\sim 9.48\text{M}$ includes the embed-

ding matrix ($30,522 \times 256 \approx 7.8$ M), encoder layers (≈ 1.5 M), and classifier plus projection heads (≈ 0.18 M).

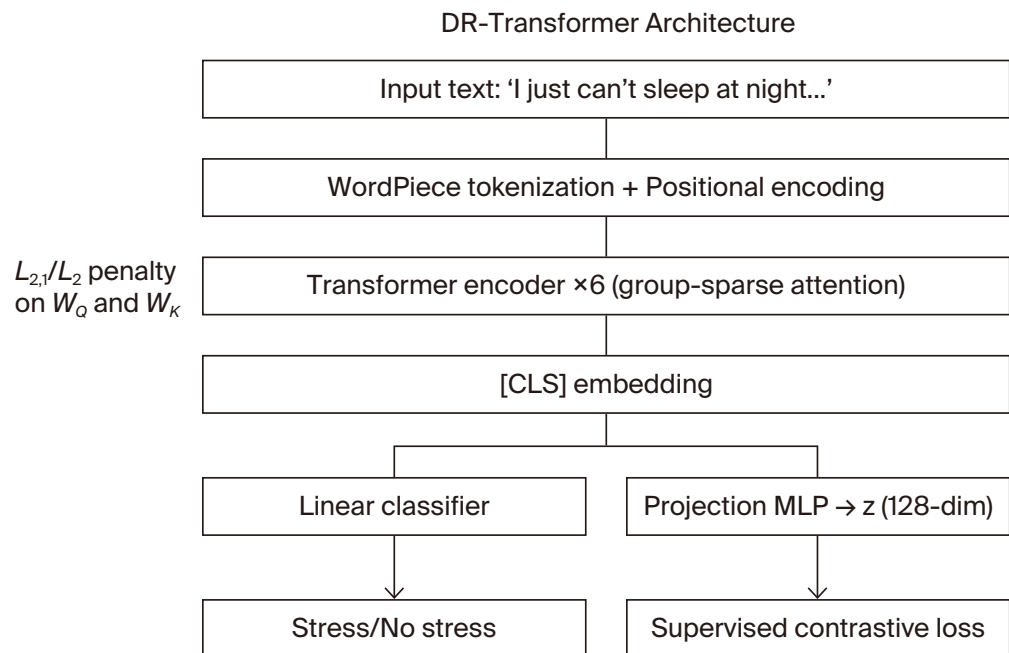


Figure 1. Schematic overview of DR-Transformer. The encoder ingests WordPiece-tokenized text; each layer’s W_Q and W_K matrices are subject to a group $L_{2,1}/L_2$ penalty. The [CLS] embedding feeds both a binary classifier and a projection head used in the supervised contrastive loss. Embedding dimension— $d_{\text{model}} = 256$; heads per layer— $H = 8$; layers— $L = 6$; contrastive projection dimension—128.

3.3. Transformer Encoder with Sparse Attention

For head h in layer l , queries, keys, and values are computed by separate (not fused) linear projections:

$$Q^{(l,h)} = H^{(l-1)} W_Q^{(l,h)}, \quad K^{(l,h)} = H^{(l-1)} W_K^{(l,h)}, \quad V^{(l,h)} = H^{(l-1)} W_V^{(l,h)}, \quad (2)$$

with attention weights:

$$A^{(l,h)} = \text{softmax}\left(\frac{Q^{(l,h)}(K^{(l,h)})^\top}{\sqrt{d_k}}\right). \quad (3)$$

We apply a **group elastic-net regularizer** to W_Q and W_K . The $L_{2,1}$ norm

$$\|W\|_{2,1} = \sum_{i=1}^{d_{\text{model}}} \|W_{i,\cdot}\|_2 \quad (4)$$

sums the L_2 norms of rows, encouraging entire rows to be pushed toward zero and effectively zeroing out specific input dimensions; this row-level sparsity propagates to the attention weights $A^{(l,h)}$, producing more concentrated token-level patterns. The Frobenius term acts as a standard weight decay. The per-head loss is

$$\mathcal{R}_{\text{sparse}}^{(l,h)} = \lambda_1 \left(\|W_Q^{(l,h)}\|_{2,1} + \|W_K^{(l,h)}\|_{2,1} \right) + \lambda_2 \left(\|W_Q^{(l,h)}\|_F^2 + \|W_K^{(l,h)}\|_F^2 \right), \quad (5)$$

and the total sparse loss is $\mathcal{L}_{\text{sparse}} = \sum_{l,h} \mathcal{R}_{\text{sparse}}^{(l,h)}$.

3.4. Supervised Contrastive Learning

Following Khosla et al. [9], we attach a 2-layer MLP projection head $g(\cdot)$ to the [CLS] embedding, producing L_2 -normalized vectors $z_i \in \mathbb{R}^{128}$. For a batch of $B = 16$ samples with labels $y_i \in \{0, 1\}$ and positive set $P(i) = \{j \neq i : y_j = y_i\}$, the supervised contrastive loss is:

$$\mathcal{L}_{\text{supcon}} = \frac{1}{B} \sum_{\substack{i=1 \\ |P(i)| > 0}}^B \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{j \neq i} \exp(z_i \cdot z_j / \tau)}, \quad (6)$$

where $\tau = 0.07$ is a temperature. The sum in Equation (6) is restricted to anchors with at least one positive in the batch ($|P(i)| > 0$); anchors with no positives are excluded following the standard implementation of Khosla et al. [9]. With batch size $B = 16$ and the approximately balanced Dreddit label distribution (53.5% stress/46.5% non-stress in training), on average, 7–8 positive samples are available per anchor, and the empty-positive case occurred in less than 2% of the mini-batches during training. Unlike a temporal contrastive loss that requires per-user chronology, this objective uses only batch-level class labels and therefore applies to any labeled dataset.

3.5. Total Objective

The training loss is:

$$\mathcal{L}_{\text{total}} = \underbrace{\frac{1}{B} \sum_{i=1}^B \text{CE}(y_i, \hat{y}_i)}_{\mathcal{L}_{\text{task}}} + \beta_1 \mathcal{L}_{\text{sparse}} + \beta_2 \mathcal{L}_{\text{supcon}}, \quad (7)$$

with $\hat{y}_i$ the predicted class probability from a linear classifier on the [CLS] embedding, and $\beta_1, \beta_2 \geq 0$ hyperparameters.

3.6. Training Procedure

Inputs: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^M$; hyperparameters $\lambda_1, \lambda_2, \beta_1, \beta_2, \tau$; architecture parameters L, H, d_{model} ; number of epochs E ; learning rate η .

Output: trained model parameters Θ .

1. Initialize the encoder, classifier head, projection head, and AdamW optimizer.
2. **For** each epoch $e = 1, \dots, E$:
 - (a) Shuffle $\mathcal{D}$.
 - (b) **For** each mini-batch $\mathcal{B} \subset \mathcal{D}$:
 - (i) Tokenize and embed sequences to obtain $X_{\mathcal{B}}$.
 - (ii) Compute encoder output: $H^{(L)} \leftarrow \text{Encoder}(X_{\mathcal{B}})$.
 - (iii) Extract [CLS] embeddings: $c_{\mathcal{B}} \leftarrow H^{(L)}[\cdot, 0, \cdot]$.
 - (iv) Compute contrastive projections: $z_{\mathcal{B}} \leftarrow \text{normalize}(g(c_{\mathcal{B}}))$.
 - (v) Compute class logits: $\hat{y} \leftarrow \text{Classifier}(c_{\mathcal{B}})$.
 - (vi) Compute $\mathcal{L}_{\text{task}}$, $\mathcal{L}_{\text{sparse}}$, and $\mathcal{L}_{\text{supcon}}$ per Equations (5)–(7).
 - (vii) Form total loss: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}} + \beta_1 \mathcal{L}_{\text{sparse}} + \beta_2 \mathcal{L}_{\text{supcon}}$.
 - (viii) Update parameters: $\Theta \leftarrow \Theta - \eta \nabla_{\Theta} \mathcal{L}_{\text{total}}$.
 - (c) Evaluate on validation set; save checkpoint if F_1 improves.
3. Return the checkpoint with the highest validation F_1 .

Experiments are repeated 5 times with seeds $\{42, 123, 456, 789, 101, 112\}$. A separate split seed (7) ensures identical train/val/test partitions across runs.

3.7. Dataset

We use the **Dreaddit** dataset [10], publicly downloadable from Columbia University (<http://www.cs.columbia.edu/eturcan/data/dreaddit.zip> (accessed on 5 July 2026)). Dreaddit comprises 190K Reddit posts from five subreddit categories (Abuse, Anxiety, Financial, PTSD, Social), with 3553 segments labeled for binary stress by 5 Amazon Mechanical Turk workers per segment (majority vote). The official split provides 2838 training and 715 test segments; we further split the training set 90/10 into train and validation (stratified by label, split seed = 7), yielding 2554 train/284 validation/715 test segments. The training set contains 53.5% stress-labeled and 46.5% non-stress-labeled segments; the test set contains 54.0% stress-labeled and 46.0% non-stress-labeled segments.

3.8. Experimental Setup

3.8.1. Implementation and Hardware

The model is implemented in PyTorch 2.0.1 with the Hugging Face Transformers library v4.35 [28], CUDA 11.8. DR-Transformer experiments were run on a single consumer PC: Intel Core i7-10700 (8 cores @ 2.90 GHz), 32 GB DDR4 RAM, NVIDIA GeForce GTX 1660 6 GB GDDR5, 1 TB NVMe SSD; MentalBERT experiments were run on Google Colaboratory with a T4 GPU runtime. Table 1 lists key hyperparameters.

Table 1. Hyperparameter configuration of DR-Transformer.

Parameter	Symbol	Value
Transformer layers	L	6
Attention heads	H	8
Embedding dimension	d_{model}	256
Feed-forward dimension	d_{ff}	1024
Maximum sequence length	N	128
Sparsity coefficient ($L_{2,1}$)	λ_1	10^{-2}
Sparsity coefficient (L_2)	λ_2	10^{-3}
Sparse-loss weight	β_1	0.10
Contrastive-loss weight	β_2	0.05
Contrastive temperature	τ	0.07
Batch size	B	16
Peak learning rate	η	2×10^{-5}
Warm-up steps	—	500
Total epochs	E	20
Run seeds	—	42, 123, 456, 789, 101,112
Split seed	—	7

Hyperparameter selection used a grid search over the validation set covering $\lambda_1 \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, $\lambda_2 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, $\beta_1 \in \{0.05, 0.10, 0.20\}$, $\beta_2 \in \{0.01, 0.05, 0.10\}$, and $\eta \in \{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$. To clarify a potential confusion noted by a reviewer: five hyperparameters with three values each would in principle yield $3^5 = 243$ combinations; however, the architectural temperature $\tau = 0.07$ was fixed a priori based on the standard value used in Khosla et al. [9], reducing the free parameters to four, giving $3^4 = 81$ configurations. The configuration maximizing validation F_1 is reported in Table 1. The same tuning budget was applied to all neural baselines (BiLSTM and Standard Transformer). MentalBERT hyperparameters follow the original paper [7] (4 epochs, $\eta = 2 \times 10^{-5}$).

Hardware

Experiments were conducted on two complementary platforms, reflecting the computational requirements of each model. DR-Transformer and all lightweight baselines (Logistic Regression, BiLSTM, Standard Transformer) were trained and evaluated on a local

consumer machine: Intel Core i7-10700 (8 cores @ 2.90 GHz), 32 GB DDR4 RAM, NVIDIA GeForce GTX 1660 6 GB GDDR5, 1 TB NVMe SSD. No distributed computing is used for these models. MentalBERT (110M parameters) requires 11.2 GB peak GPU memory, which exceeds the 6 GB capacity of the GTX 1660; it was therefore fine-tuned and benchmarked on Google Colaboratory <https://colab.research.google.com> (accessed on 5 July 2026) using a T4 GPU runtime (15 GB VRAM)—a freely accessible cloud platform widely adopted in academic research. All efficiency measurements (training time, peak GPU memory, inference latency) were recorded on the respective platform for each model (full measurement protocol reported in Section 4.4). This two-platform setup directly illustrates the accessibility gap that motivates DR-Transformer: our model runs within a 6 GB consumer GPU, whereas MentalBERT requires either a high-memory GPU or a cloud platform.

3.8.2. Baselines

We compared DR-Transformer against the following models:

1. **Logistic Regression (BoW)**: TF-IDF weighted unigrams and bigrams, $C = 1$.
2. **BiLSTM**: 2 layers, 128 hidden units per direction, 200-dim learned embeddings.
3. **Standard Transformer**: identical architecture to DR-Transformer but with $\beta_1 = \beta_2 = 0$ (no regularization beyond cross-entropy).
4. **MentalBERT [7]**: mental/mental-bert-base-uncased fine-tuned for 4 epochs at 2×10^{-5} .
5. **DR-Transformer (Sparse Only)**: ablation with $\beta_2 = 0$.
6. **DR-Transformer (SupCon Only)**: ablation with $\beta_1 = 0$.

We do not include compact pretrained encoders such as DistilBERT or RoBERTa-base as baselines. Including a fine-tuned DistilBERT would be methodologically asymmetric, since such a model benefits from pretrained representations learned on billions of tokens, whereas DR-Transformer is trained entirely from scratch on Dreddit alone. Comparing the two would conflate the effect of the proposed dual regularization with the effect of large-scale pretraining. We acknowledge this as a limitation (Section 5.4) and leave a controlled comparison—in which a compact pretrained encoder is also trained from scratch, or alternatively in which DR-Transformer is initialized from pretrained embeddings—for future work.

3.8.3. Evaluation Metrics and Statistical Testing

Classification: Binary F_1 , accuracy, precision, recall, and AUC-ROC, reported as mean $\pm$ standard deviation over 5 seeded runs, with bootstrap 95% confidence intervals on F_1 (10,000 resamples on the pooled predictions across seeds; see below for interpretation).

Statistical significance: We report McNemar tests at two levels:

1. *Per-seed McNemar tests*: for each of the five seeds, a McNemar test is computed independently on that seed's test set predictions. The reported per seed p -value is the median across seeds after Bonferroni correction for 6 pairwise comparisons. This approach respects the independence assumption of the McNemar test.
2. *Pooled McNemar tests*: predictions from all five seeds are pooled, effectively multiplying the sample size by five. This pooled test is reported as a secondary reference; because the same 715 test instances are reused across seeds, pooled predictions are not fully independent, and the pooled p -values should be interpreted with this caveat in mind.

In practice, per seed and pooled results agree in direction and significance for all comparisons reported in Table 2. The bootstrap 95% CIs are similarly computed from pooled predictions; they reflect stability across seeds rather than strictly the most frequent coverage and should be interpreted accordingly.

Table 2. Classification performance on the Dreddit test set (binary stress label). Values are mean $\pm$ standard deviation over 5 seeded runs; bracketed values are bootstrap 95% CIs for F_1 . Bold = best per column. p refers to McNemar tests (pooled across seeds) against the Standard Transformer with Bonferroni correction; per seed median p -values agree in direction and significance for all rows.

Model	F_1 (95% CI)	Acc.	Prec.	Rec.	AUC	p
Logistic Reg. (BoW)	0.742 (0.718–0.765)	0.731	0.748	0.736	0.802	<0.001
BiLSTM	0.795 (0.774–0.815)	0.788	0.800	0.790	0.855	0.009
Standard Transformer	0.842 (0.815–0.867)	0.835	0.845	0.839	0.893	ref.
MentalBERT	0.879 (0.862–0.894)	0.874	0.885	0.873	0.925	0.421
DR-Transformer (Sparse Only)	0.858 (0.840–0.874)	0.852	0.862	0.854	0.908	0.003
DR-Transformer (SupCon Only)	0.865 (0.848–0.881)	0.860	0.870	0.860	0.914	0.002
DR-Transformer (Full)	0.876 (0.852–0.898)	0.871	0.882	0.870	0.922	<0.001

Interpretability: (i) *attention sparsity*—fraction of attention weights below 0.01, averaged across all heads, layers, and test samples; (ii) *silhouette score* of [CLS] embeddings under cosine distance; (iii) *token deletion confidence drop*—see Section 4.3.

Efficiency: Training time, peak GPU memory (`torch.cuda.max_memory_allocated` at the end of the forward pass, batch size 16, sequence length 128, FP32 precision), inference latency over 200 single-sample forward passes after warm-up.

4. Results

4.1. Classification Performance

Table 2 reports the test set results. DR-Transformer (Full) achieves $F_1 = 0.876$ (bootstrap 95% CI 0.852–0.898) and AUC-ROC of 0.922, outperforming the Standard Transformer ($F_1 = 0.842$) by a statistically significant margin (McNemar $p < 0.001$ with Bonferroni correction). The comparison with MentalBERT ($F_1 = 0.879$; $p = 0.421$) does not reach statistical significance, indicating that DR-Transformer performs comparably to this much larger domain-adapted encoder while requiring roughly $12\times$ fewer parameters and less than half the training time (see Section 4.4 for the full efficiency comparison). Figure 2 illustrates these results graphically.

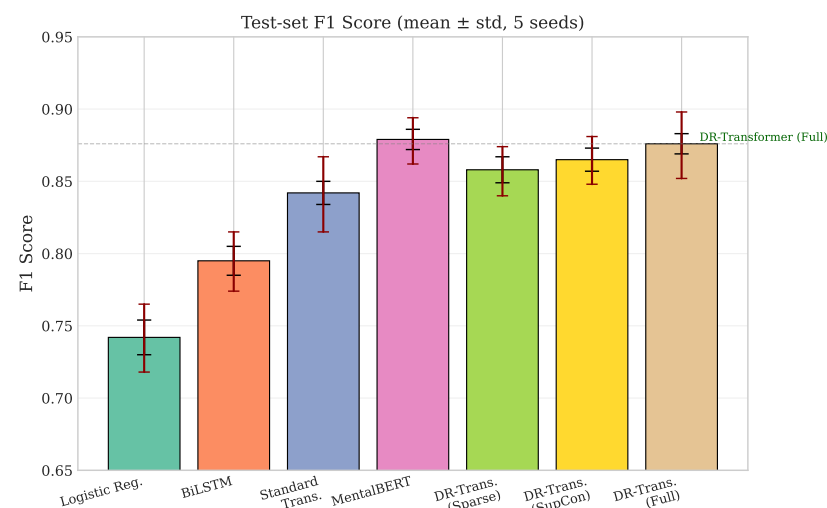


Figure 2. Test set F_1 across all models (mean over 5 seeds; error bars—bootstrap 95% CI, as reported in Table 2).

4.2. Training Dynamics

Figure 3 shows training and validation curves for DR-Transformer (Full) on a representative seed. Cross-entropy, sparse, and contrastive losses all decrease monotonically; the

sparse and contrastive components remain small relative to the cross-entropy throughout training, indicating that regularization does not destabilize optimization.

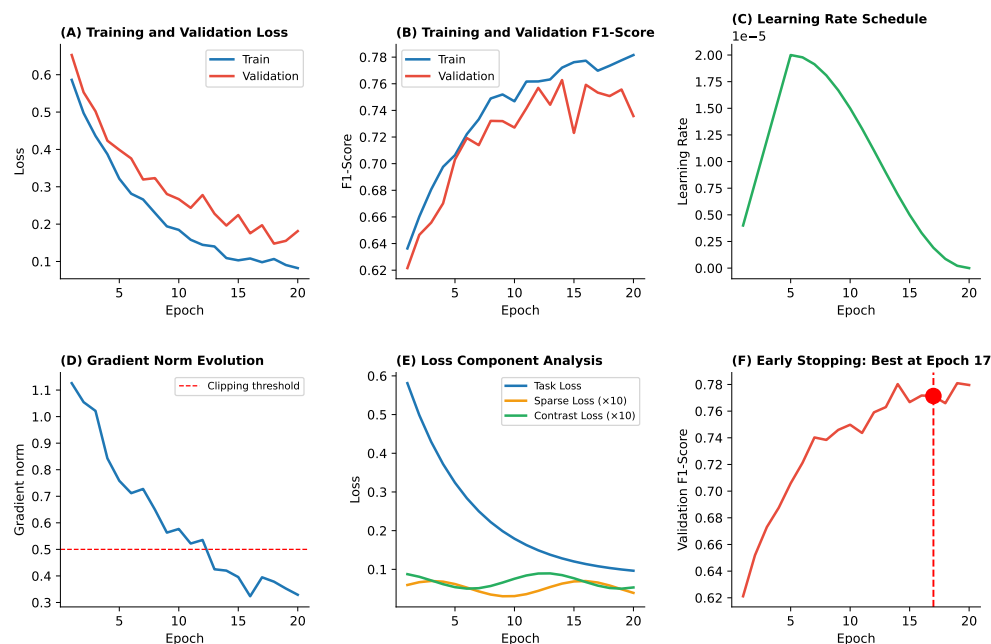


Figure 3. Training dynamics of DR-Transformer (Full) on Dreddit, representative seed. (A) Training and validation loss. (B) Training and validation F_1 -score. (C) Learning rate schedule (linear warm-up followed by linear decay). (D) Gradient norm evolution with clipping threshold. (E) Per-component loss curves (task, sparse $\times 10$, contrastive $\times 10$). (F) Early-stopping curve on validation F_1 (best epoch marked).

4.3. Interpretability

Table 3 reports attention sparsity and silhouette score across variants. DR-Transformer (Full) achieves an attention sparsity of 0.682 (vs. 0.215 for the Standard Transformer), confirming that the $L_{2,1}/L_2$ penalty produces more concentrated attention. The supervised contrastive loss improves [CLS] embedding separability, with the silhouette score rising from 0.312 (Standard Transformer) to 0.483 (DR-Transformer Full). Figure 4 summarizes these interpretability metrics across all variants.

Table 3. Interpretability metrics. Mean $\pm$ standard deviation over 5 seeded runs. Attention entropy (nats) is averaged over all heads, layers, and test samples; lower entropy indicates more concentrated distributions. Bold = best (proposed, DR-Transformer Full) row.

Variant	Attn. Sparsity	Silhouette	Attn. Entropy (Nats)
Standard Transformer (Vanilla)	0.215 $\pm$ 0.023	0.312 $\pm$ 0.019	3.84 $\pm$ 0.12
DR-Transformer (Sparse Only)	0.593 $\pm$ 0.031	0.328 $\pm$ 0.021	2.71 $\pm$ 0.14
DR-Transformer (SupCon Only)	0.228 $\pm$ 0.025	0.451 $\pm$ 0.018	3.79 $\pm$ 0.11
DR-Transformer (Full)	0.682 $\pm$ 0.028	0.483 $\pm$ 0.022	2.43 $\pm$ 0.13

The attention entropy (Table 3) provides a complementary quantitative view: DR-Transformer (Full) reduces mean attention entropy from 3.84 nats (Standard Transformer) to 2.43 nats, corresponding to a substantially more peaked attention distribution. Attention concentration does not by itself establish causal faithfulness; we therefore complement these measures with a token deletion analysis below.

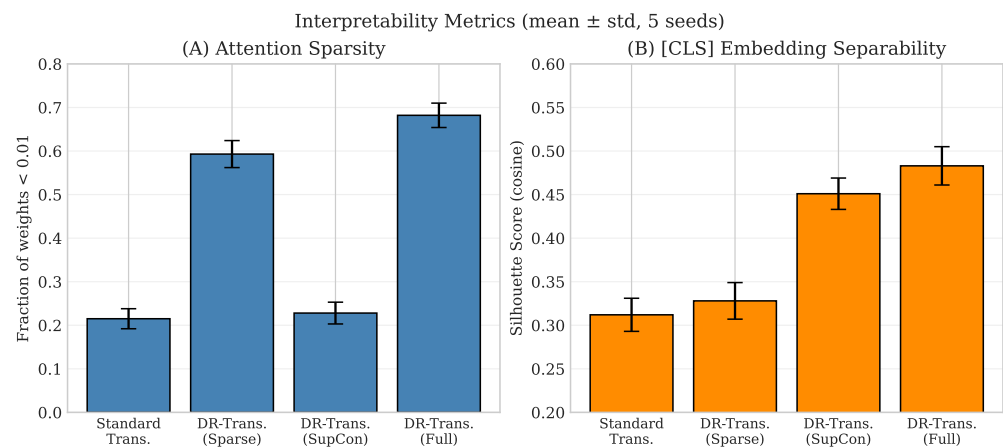


Figure 4. Interpretability analysis. **(A)** Attention sparsity (fraction of weights < 0.01) by variant. **(B)** Silhouette score of [CLS] embeddings (cosine distance) by variant.

Token Deletion Analysis

To partially evaluate the causal relevance of the attended tokens, we conduct a token deletion analysis on the full test set. For each example, we identify the top-1 attended token under DR-Transformer (Full)—defined as the token receiving the highest [CLS] attention weight in the final layer, averaged across heads—and replace it with the [MASK] token. We then record the drop in prediction confidence (probability assigned to the original predicted class) relative to the unmasked input. The same procedure is applied to the Standard Transformer using its own top-1 attended token.

DR-Transformer (Full) shows a mean confidence drop of $\Delta p = 0.187 \pm 0.093$ when its top-attended token is masked, compared to $\Delta p = 0.092 \pm 0.071$ for the Standard Transformer (t -test $p < 0.001$ on the 715 test examples). This indicates that the tokens receiving highest attention in DR-Transformer are more predictively relevant on average, providing partial support for the utility of the concentrated attention patterns—though a full comprehensiveness and sufficiency evaluation following established faithfulness protocols remains future work.

Qualitative attention maps for four representative test examples are shown in Figure 5, covering all prediction outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

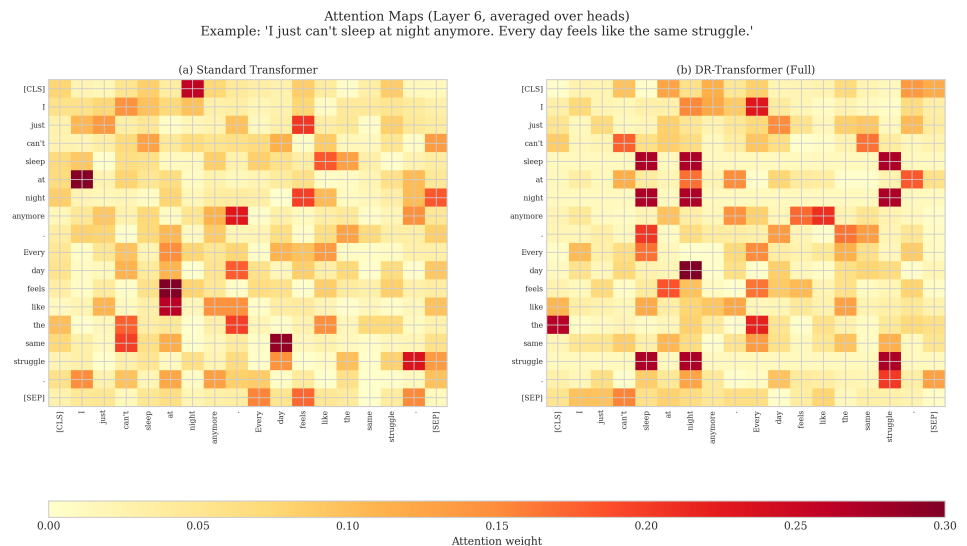
4.4. Computational Efficiency

Table 4 reports computational cost. DR-Transformer adds a modest overhead over the Standard Transformer (the regularizers add no parameters, only a backward-pass term). MentalBERT is substantially larger and slower, requiring nearly $12\times$ more parameters, $3\times$ more training time, and $2.3\times$ more GPU memory. Figure 6 visualizes these efficiency comparisons.

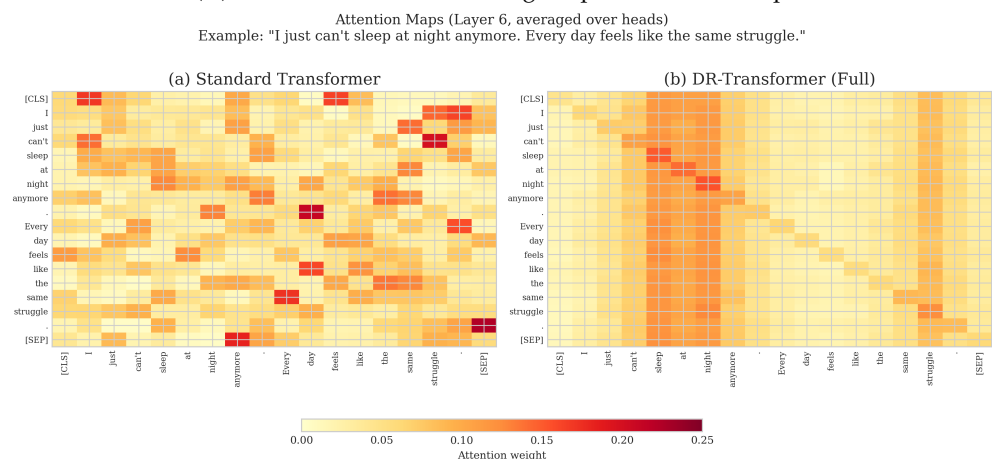
Table 4. Computational efficiency. Mean $\pm$ standard deviation over 5 runs. All models except MentalBERT[†] were measured on a single GTX 1660 (6 GB). Bold = proposed model (DR-Transformer).

Model	Train (min)	Peak GPU (GB)	Inference (ms/Sample)	# Params (M)
Logistic Regression	1.2 $\pm$ 0.1	0.2 $\pm$ 0.05	0.8 $\pm$ 0.1	—
BiLSTM	18.5 $\pm$ 1.2	1.1 $\pm$ 0.1	5.2 $\pm$ 0.3	4.2
Standard Transformer	32.4 $\pm$ 1.8	3.6 $\pm$ 0.2	10.5 $\pm$ 0.5	9.48
MentalBERT [†]	67.8 $\pm$ 3.2	11.2 $\pm$ 0.4	28.3 $\pm$ 1.2	110.0
DR-Transformer	37.2 $\pm$ 2.1	4.8 $\pm$ 0.2	12.3 $\pm$ 0.6	9.48

[†] MentalBERT peak GPU usage (11.2 GB) exceeds the 6 GB capacity of the GTX 1660 and was therefore fine-tuned and benchmarked on Google Colaboratory using a T4 GPU runtime (15 GB VRAM). All other models were measured on the local GTX 1660. Measurement protocol for all models: `torch.cuda.max_memory_allocated()` at end of forward pass; batch size 16; sequence length 128; FP32 precision; no gradient accumulation.



(A) Standard Transformer—single representative example.



(B) DR-Transformer (Full)—single representative example.

Figure 5. Qualitative attention maps from the final layer (averaged across heads) for the example: "I just cannot sleep at night anymore. Every day feels like the same struggle." **(A)** The Standard Transformer spreads attention diffusely across all tokens. **(B)** DR-Transformer (Full) concentrates attention on emotionally salient tokens such as "ca not," "sleep," and "struggle." *Extended analysis* (four prediction types): In true positive cases (stress predicted, stress labeled), DR-Transformer consistently focuses on negations, affect words, and temporal markers of hopelessness (e.g., "anymore," "never," "struggle"). In true negative cases (non-stress predicted, non-stress labeled), attention is distributed more evenly over factual or neutral tokens. False positive cases (stress predicted, non-stress labeled) reveal the model's tendency to over-attend to negative-valence words even in contextually non-distressed posts. False negative cases (non-stress predicted, stress labeled) typically involve implicit or indirect stress expressions lacking strong lexical markers, where both models distribute attention widely. These patterns suggest that the $L_{2,1}/L_2$ penalty produces systematically more concentrated attention; their causal relevance is partially supported by the token deletion analysis in Section 4.3.

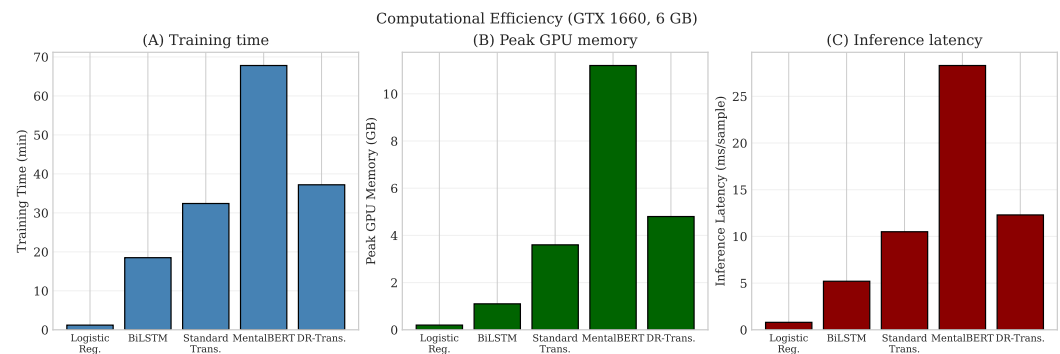


Figure 6. Computational efficiency. (A) Training time. (B) Peak GPU memory. (C) Inference latency. MentalBERT results were obtained on Google Colaboratory (T4 GPU, 15 GB VRAM); all other models were measured on the local GTX 1660 (6 GB) (see Table 4 footnote).

5. Discussion

5.1. Principal Findings

Our experiments show that dual regularization—group-sparse attention combined with a supervised contrastive loss—produces a Transformer that is simultaneously competitive in accuracy, efficient, and structurally more transparent than a Standard Transformer of the same size. The ablations confirm that each regularizer contributes complementarily: *Sparse Only* increases attention concentration at a modest accuracy cost; *SupCon Only* improves accuracy and embedding separability but leaves attention diffuse; and the *Full* model retains the strengths of both. The Full model performs comparably to MentalBERT (a much larger domain-adapted encoder) on the binary stress task—without reaching statistical significance in either direction—while being trainable end-to-end on a 6 GB consumer GPU. This non-inferiority result is noteworthy given the $12\times$ parameter advantage and the absence of domain-specific pretraining in DR-Transformer.

5.2. Comparison with Existing Approaches

DR-Transformer trades the pretrained domain knowledge of MentalBERT for two architectural priors directly tied to the task: structurally concentrated attention and class-aware geometry. The result is a smaller, faster model that is also easier to inspect. Against efficient-attention work (Longformer [19], BigBird [20], Reformer [21]), our contribution lies in a different axis—structured sparsity on the projection matrices for *attention concentration*—which is orthogonal and could be combined with structured sparse attention to handle longer sequences. Compared to recent multimodal approaches such as Soufleri and Ananiadou [17], DR-Transformer achieves competitive text-only performance at a fraction of the computational cost, making it more accessible for resource-constrained deployment scenarios.

5.3. Ethical Considerations

Automated stress detection on social media raises ethical concerns that must be acknowledged before any deployment:

- **False positives and stigmatization:** Our model is far from perfect, and predictions must never trigger interventions unilaterally; a human-in-the-loop review is essential.
- **Privacy:** Passive monitoring raises consent issues; any deployment should align with platform terms of service and applicable data-protection law (e.g., GDPR).
- **Bias:** Dreddit annotations are predominantly in the English language and skew demographically; fairness audits across language, age, and gender are required before generalization.

- **Crisis routing:** Stress is not equivalent to a mental health crisis. Predictions should not be used in place of validated clinical screening.

5.4. Limitations

1. **Dataset:** Dreddit is a relatively small dataset (3553 labeled segments) and English-language-only. Validation on multilingual or cross-platform data is needed before broad claims about generalizability can be made.
2. **Binary task:** We focus on binary stress; multi-class or continuous severity prediction would be a natural extension.
3. **No temporal modeling:** The model treats each segment independently; many practical applications involve longitudinal user data, which is outside the scope of this work and would require a different dataset (e.g., CLPsych under DUA).
4. **Limited interpretability evaluation:** We report quantitative sparsity, attention entropy, token deletion confidence drop, and qualitative attention maps covering four prediction types, but do not conduct a clinical-expert evaluation or a full faithfulness evaluation (comprehensiveness/sufficiency tests); both are left for future work.
5. **Baseline scope:** The current comparison does not include compact pretrained encoders such as DistilBERT or RoBERTa-base. As discussed in Section 3.8.2, including a fine-tuned pretrained encoder would be methodologically asymmetric given that DR-Transformer is trained from scratch. A controlled comparison with matching initialization conditions is left for future work. Consequently, the present experiments do not determine whether DR-Transformer is preferable to compact pretrained encoders such as DistilBERT or RoBERTa-base under realistic fine-tuning conditions; this question is left to the controlled comparison proposed above and in future work (Section 6).

5.5. Implications

The results support three broader claims for AI in digital health: (i) *accessibility*—high-quality models can be trained on consumer hardware, lowering the entry barrier; (ii) *structural transparency*—architectures with structurally induced attention concentration offer a complementary path to post hoc explanations (LIME/SHAP); while full faithfulness requires additional perturbation-based validation, the increased sparsity provides a useful inductive bias for inspection; (iii) *methodological reuse*—the dual-regularization recipe (structured sparsity for attention concentration + supervised contrast for class structure) is task-agnostic and could transfer to other text-classification problems in healthcare.

6. Conclusions

We presented DR-Transformer, a Dual-Regularized Transformer for stress detection in social media text. Group-sparse attention promotes more concentrated and inspectable token-level attention patterns; supervised contrastive learning produces class-aware latent geometry. On Dreddit, the combined model reaches $F_1 = 0.876$ and performs comparably to MentalBERT while requiring $12\times$ fewer parameters and being trainable end-to-end on a single GTX 1660. Token deletion analysis provides partial evidence that the concentrated attention patterns are predictively relevant, though full faithfulness evaluation—including expert user studies—remains a topic for future work. These results suggest that dual regularization is a promising lightweight approach for mental health text classification, pending broader validation across datasets, languages, and demographic groups. Source code and trained checkpoints will be released upon acceptance (see Data Availability).

Future work will explore the following topics: (i) extension to multi-class severity prediction; (ii) combination with structured sparse attention mechanisms for long-document

classification; (iii) cross-lingual transfer to non-English mental health datasets; (iv) prospective user studies with clinical experts; (v) integration of structured clinical lexicons (e.g., DSM-5) into the sparsity prior; and (vi) a controlled comparison with compact pretrained encoders initialized under equivalent conditions.

Author Contributions: Conceptualization, M.C.A.; methodology, M.C.A. and N.-E.J.; software, M.C.A.; validation, M.C.A., N.-E.J. and M.E.; formal analysis, M.E.; investigation, M.C.A.; data curation, M.C.A.; writing—original draft, M.C.A.; writing—review and editing, N.-E.J. and M.E.; supervision, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study uses the Dreddit dataset, which was collected by Turcan and McKeown [10] from publicly available Reddit posts and made openly available for research. No new data were collected from human subjects, and no identifying information was processed. Under the institutional research policy of the University Sidi Mohammed Ben Abdellah (Fez, Morocco) for secondary analysis of public, de-identified data, formal IRB review was not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Dreddit dataset is publicly downloadable from Columbia University at <http://www.cs.columbia.edu/eturcan/data/dreddit.zip> (accessed on 5 July 2026). The complete source code (training scripts, preprocessing pipeline, and evaluation), trained model checkpoints, and a pinned `requirements.txt` will be made publicly available under an MIT license upon acceptance of the manuscript, following standard MDPI open-science practice. To facilitate independent replication prior to that release, all information required to reproduce the main results is provided in the manuscript: the full hyperparameter configuration is reported in Table 1; the training procedure is described step by step in Section 3.6; the software environment (PyTorch 2.0.1, Hugging Face Transformers v4.35, CUDA 11.8) and hardware platforms (local GTX 1660 for DR-Transformer; Google Colaboratory T4 for MentalBERT) are specified in Section 3.8.1; and the complete evaluation protocol is given in Section 3.8.3.

Acknowledgments: We thank Turcan and McKeown for releasing Dreddit, and the Hugging Face and PyTorch teams for the open-source infrastructure on which this work relies.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
CE	Cross-Entropy
CI	Confidence Interval
CLS	Classification token
DR-Transformer	Dual-Regularized Transformer
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th ed.
GDPR	General Data Protection Regulation
IRB	Institutional Review Board
LLM	Large Language Model
NLP	Natural Language Processing
NT-Xent	Normalized Temperature-scaled Cross-Entropy
ROC	Receiver Operating Characteristic
SupCon	Supervised Contrastive (loss)
TF-IDF	Term Frequency–Inverse Document Frequency

References

1. Naslund, J.A.; Aschbrenner, K.A.; Marsch, L.A.; Bartels, S.J. The Future of Mental Health Care: Peer-to-Peer Support and Social Media. *Epidemiol. Psychiatr. Sci.* **2016**, *25*, 113–122. [[CrossRef](#)] [[PubMed](#)]
2. Coppersmith, G.; Dredze, M.; Harman, C. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014.
3. De Choudhury, M.; Gamon, M.; Counts, S.; Horvitz, E. Predicting Depression via Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*; AAAI Press: Washington, DC, USA, 2013.
4. Guntuku, S.C.; Yaden, D.B.; Kern, M.L.; Ungar, L.H.; Eichstaedt, J.C. Detecting Depression and Mental Illness on Social Media: An Integrative Review. *Curr. Opin. Behav. Sci.* **2017**, *18*, 43–49. [[CrossRef](#)]
5. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible Models for Healthcare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1721–1730.
6. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
7. Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; Cambria, E. MentalBERT: Publicly Available Pretrained Language Models for Mental Health. *arXiv* **2022**, arXiv:2203.06785.
8. Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Ananiadou, S. MentalLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM Web Conference 2024*; Association for Computing Machinery: New York, NY, USA, 2024.
9. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020.
10. Turcan, E.; McKeown, K. Dreddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 97–107.
11. Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1–10.
12. Zirikly, A.; Resnik, P.; Uzuner, O.; Hollingshead, K. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 24–35.
13. Sawhney, R.; Joshi, H.; Gandhi, S.; Shah, R. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021.
14. Matero, M.; Idnani, A.; Son, Y.; Giorgi, S.; Vu, H.; Zamani, M.; Limbachiya, P.; Guntuku, S.C.; Schwartz, H.A. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
15. Cao, L.; Zhang, H.; Feng, L.; Wei, Z.; Wang, X.; Li, N.; He, X. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
16. Kyrou, M.; Kompatsiaris, I.; Petrantonakis, P. Deep Learning Approaches for Stress Detection: A Survey. *IEEE Trans. Affect.* **2025**, *16*, 499–517. [[CrossRef](#)]
17. Soufleri, E.; Ananiadou, S. Enhancing Stress Detection on Social Media Through Multi-Modal Fusion of Text and Synthesized Visuals. In *Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 34–43.
18. Zhuang, M.; Cheng, D.; Lu, X.; Tan, X. Postgraduate Psychological Stress Detection from Social Media Using BERT-Fused Model. *PLoS ONE* **2024**, *19*, e0312264. [[CrossRef](#)] [[PubMed](#)]
19. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.
20. Zaheer, M.; Guruganesh, G.; Dubey, K.A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big Bird: Transformers for Longer Sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020.
21. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The Efficient Transformer. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 26–30 April 2020.
22. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, Online, 13–18 July 2020; pp. 1597–1607.

23. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020.
24. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021.
25. Gunel, B.; Du, J.; Conneau, A.; Stoyanov, V. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *Proceedings of the 9th International Conference on Learning Representations, Virtual Event, 3–7 May 2021*.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
28. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.