IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MTL-DAPT: A Domain-Adapted Multi-Task Learning Framework for Joint Depression, Anxiety, and Suicidality Detection from Social Media

Yukta Chourasiya[1], Vinita Shrivastava[2]

[1,2] Department of Computer Science and Engineering, Oriental Institute of Science and Technology, Bhopal, India

ABSTRACT—Current computational methods for detecting mental health status from social media generally frame the problem as a binary classification of individual conditions, overlooking the clinical reality that depression, anxiety and suicidality are often comorbid at high rates. Here we present MTL-DAPT, a Multi-Task Learning framework based on a Domain-Adaptively Pre-Trained RoBERTa-base encoder that jointly predicts all the three conditions from a single shared representation. First, the encoder is adapted to mental health language by masked language modeling on 50,000 Reddit posts sampled from the Reddit Mental Health Dataset (RMHD, 2020–2022) and then fine-tuned end-to-end with three independent binary task heads under a class-weighted multi-task loss. On a held-out, stratified test set of 22,500 posts, the proposed model achieves a macro-averaged F1 of 0.7970 and a macro-averaged AUC-ROC of 0.8966 across the three tasks: 0.7261 F1 / 0.8284 AUC for depression, 0.8732 F1 / 0.9627 AUC for anxiety, and 0.7918 F1 / 0.8988 AUC for suicidality. The robustness of the result is verified using three-fold cross-validation (mean F1 = 0.7941, standard deviation = 0.0008). This performance is further highlighted by two controlled baselines . A single-task BERT-base classifier trained only on depression achieves an F1 of just 0.7120 . An ablation of three separate DAPT-RoBERTa classifiers (one per condition, identical encoder, no shared training) achieves a macro F1 of 0.7979 , statistically indistinguishable from the jointly trained model (−0.0009 F1). Thus the domain adaptation step alone is responsible for the bulk of the improvement over the plain-BERT baseline, and joint multi-task training offers single-task performance from a single unified model rather than three independent ones, cutting the number of models trained by two-thirds at no measurable accuracy cost. To the authors' knowledge, no previous work has been published on this dataset that has combined domain-adaptive pre-training with joint prediction of these three specific conditions, nor has reported the full precision, recall, F1 and AUC breakdown per task, or a cross-validated stability estimate as presented in this paper.

Keywords—Mental health detection, multi-task learning, domain adaptive pre-training, RoBERTa, depression, anxiety, suicidality, social media, class imbalance, natural language processing.

I. INTRODUCTION

THE phrase "digital psychiatry" entered clinical discourse around 2014 [1], describing an emerging practice of using passively generated digital data to supplement traditional mental health assessment. The World Health Organization estimates that over one billion individuals live with some form of mental illness, with the global ratio of mental health workers to population remaining critically low across much of the world [2]. Against this backdrop, computational detection of mental health status from social media text has attracted sustained research interest, and yet has, almost without exception, been framed as a single-condition classification problem: a model is trained to recognise depression, or anxiety, or suicidality, and its task ends there. That framing is a reasonable starting point, but it discards information that matters clinically. Depression and anxiety are comorbid in over half of diagnosed cases [3], and suicidality is strongly associated with both [4]; a system that models only one condition, or models each condition through an entirely separate pipeline, is working with an incomplete picture of the phenomenon it claims to detect.

Transformer-based language models — BERT [5] and its more robustly optimised successor RoBERTa [6] — have displaced earlier feature-engineered classifiers across most text classification benchmarks, including mental health detection. Domain Adaptive Pre-Training (DAPT) [7], which continues a transformer's masked language modelling objective on in-domain text before task-specific fine-tuning, has separately been shown to yield consistent downstream gains in specialised domains. Multi-task learning [8], meanwhile, offers a principled way to model related conditions jointly through a shared encoder, an approach previously applied to mental health text by Benton et al. [9], who showed that joint training across related disorders improved individual-task performance under limited data availability.

This paper reports the design, training, and evaluation of a framework that combines both ideas. A RoBERTa-base encoder is first adapted to the mental health domain through DAPT, then fine-tuned end-to-end with three independent binary classification heads sharing a common encoder, producing simultaneous predictions for depression, anxiety, and suicidality from a single forward pass. The contribution is deliberately empirical rather than architectural: rather than proposing a new attention mechanism, this work asks whether a well-understood domain adaptation procedure, combined with a well-understood multi-task learning pattern, and evaluated with the rigour that the field's own methodological reviews have found lacking [10], can support joint prediction of three clinically related conditions without sacrificing the per-condition performance that single-task models achieve.

This paper makes four contributions: First, we conduct systematic domain adaptive pre-training and report its isolated contribution over a non-adapted BERT baseline instead of attributing domain-specific gains to an unadapted encoder. Second, it casts the detection of depression, anxiety and suicidality as a joint multi-task problem, and provides a controlled ablation—three independent single-task classifiers using the same domain-adapted encoder—that isolates the contribution of joint training from that of domain adaptation. Third, it provides complete per-task precision, recall, F1-score, AUC-ROC and accuracy, as well as confusion-matrix based specificity, a reporting granularity that was identified as largely missing from the reviewed literature [10]. Fourth, it establishes the statistical stability of the reported result through three-fold cross-validation on a stratified sample of the Reddit Mental Health Dataset spanning the 2020–2022 period.

The remainder of this paper is organised as follows. Section II positions this work relative to existing approaches. Section III describes the dataset, the domain adaptation procedure, and the model architecture. Section IV details the experimental setup, evaluation metrics, and baseline models. Section V presents and discusses the results. Section VI concludes and outlines directions for future work.

II. RELATED WORK

Most published work on detecting mental health from social media trains a single classifier for one condition. Qasim et al. [11] used sentence transformers with n-gram features for depression severity classification on Reddit with a macro-F1 of 0.86 but focused only on depression. Wan, et al. [12] reported 90.18% accuracy on binary classification of depression across combined Reddit and Twitter data. Bhatt et al. [13] used RoBERTa to perform multiclass classification of depression, ADHD, and PTSD with F1 scores above 0.90. In each of these examples, the conditions considered are modeled without taking into account their potential co-occurrence in the same user. Karamat et al. [14] proposed a hybrid architecture using domain-specific transformer models, improving precision of anxiety detection based on clinically informed vocabulary.

In this literature, the use of domain adaptation is inconsistent. Ji et al. [15] introduced MentalBERT and MentalRoBERTa, which were pre-trained on mental health posts from Reddit and clinical notes and demonstrated consistent improvements over general-domain BERT and RoBERTa on depression and suicidality tasks, typically in the range of 1 to 3 F1 points, a gap that is independently replicated in this paper's own baseline comparison. Baydili et al. [16] employed a class-weighted feature selection approach together with domain-adapted pre-trained language models, and reported accuracy between 80% and 99% on Reddit and Twitter datasets for detecting depression and suicidal tendency, although the two conditions were treated via separate models rather than a shared architecture. Ajayi et al. [17] proposed a unified multiclass framework for ten mental health and cyberbullying categories, but domain adaptation was used inconsistently across category subsets, making it difficult to isolate the contribution.

Multi-task and ensemble formulations have been studied to a certain degree. Cao et al. [18] studied machine learning methods for depression detection, with an emphasis on multi-label formulations, and observed consistent but rarely reported improvements from joint training. Kasap et al. [19] developed an ensemble of transformer models for joint depression and emotion classification, obtaining F1 scores of 78–85% on suicidal ideation detection, but relegated anxiety and suicidality to separate pipelines instead of a shared encoder. Wan et al. [20], again used a related ensemble variant of the depression classification task, combining BERT, RoBERTa, and a clinical BERT variant using a voting mechanism, improving on any single constituent model by 2-3 accuracy points at the cost of a proportional increase in inference complexity. He et al. [21] conducted a scoping review of 78 studies on the detection of mental health risks and explicitly pointed out the lack of research on the modeling of depression, anxiety and suicidality together, which is the specific gap that this paper addresses. Konkolics Possobom [22] came to a similar finding in a systematic review of six mental health markers, noting that ensemble methods, despite usually achieving the best results, trade off interpretability and computational feasibility.

Earlier classical machine learning approaches established that behavioural and linguistic features carry predictive signal well before transformer-based methods became standard: Burdisso et al. [23] proposed an SVM-based framework for early depression detection over social media streams, and Shen and Rudzicz [24] applied Random Forest classifiers to LIWC-derived lexical features for anxiety detection on Reddit. Both report accuracy in the low-to-mid 80s, since surpassed by domain-adapted transformer approaches such as the one proposed here.

Reporting standards across this body of work remain uneven. Chancellor and De Choudhury [10] found that only 32 of 75 studies reviewed through 2018 reported the full set of minimum standards necessary to reproduce a predictive algorithm — sample size, feature count, algorithm choice, validation method, and performance metric together — and subsequent reviews [21] confirm the pattern persists. This paper reports the full set of minimum reporting standards for every task evaluated, together with cross-validated stability estimates absent from the majority of the studies cited above.

III. PROPOSED METHODOLOGY

A. Dataset and Preprocessing

The Reddit Mental Health Dataset (RMHD) [25] was used as the primary data source. RMHD is organised as monthly CSV files spanning multiple years, each containing posts scraped from a single subreddit together with the author username, post title, self-text body, and Unix creation timestamp. This study uses data from January 2020 through December 2022 — a window that captures the COVID-19 pandemic and its immediate aftermath, a period independently documented to have had substantial effects on population-level mental health [26] — drawn from four condition-specific subreddits: r/depression, r/Anxiety, r/SuicideWatch, and r/mentalhealth, the last serving as the negative (normal) class.

Across 159 monthly files spanning the three-year window, 1,471,845 raw posts were loaded. Each post's title and self-text were concatenated, lower-cased, stripped of URLs and non-alphanumeric characters beyond basic punctuation, and posts marked [removed] or [deleted] by Reddit's own moderation system were discarded. Posts shorter than ten words were excluded, on the basis that such posts rarely carry sufficient linguistic signal for reliable classification. Exact-duplicate post texts were removed to prevent data leakage from cross-posting or repeated submissions. This pipeline reduced the corpus to 1,199,014 usable posts distributed across four labels: depression (409,235; 34.1%), suicidal (347,852; 29.0%), normal (231,276; 19.3%), and anxiety (210,651; 17.6%).

Given the computational cost of training multiple transformer models across five separate experimental configurations (two baselines, an ablation, and a cross-validated main model) within a fixed compute budget, a stratified random sample of 150,000 posts was drawn from the cleaned corpus, preserving the original class proportions to within 0.2 percentage points of the full-corpus distribution. This sampled set was partitioned 70/15/15 into training (108,374 posts), validation (19,125 posts), and test (22,500 posts) splits using stratified sampling on the label column, ensuring balanced class representation in every partition.

B. Domain Adaptive Pre-Training (DAPT)

RoBERTa-base [6] was pre-trained by its original authors on general-domain English text — news, books, and web content — and, in its released form, has no particular exposure to the informal, emotionally charged language typical of mental health discussion on Reddit. Following the domain adaptation procedure introduced by Gururangan et al. [7], the encoder was subjected to continued masked language model (MLM) pre-training on 50,000 posts sampled uniformly at random (without label stratification) from the cleaned RMHD corpus, for three epochs at a per-device batch size of 8, with a masking probability of 0.15 applied to each input sequence following the standard MLM protocol. A post-hoc audit of this 50,000-post sample confirmed that all four condition labels were represented in proportions matching the source corpus to within 0.2 percentage points — depression 34.2%, suicidal 29.1%, normal 19.1%, and anxiety 17.6% — confirming that the resulting encoder, referred to hereafter as DAPT-RoBERTa, was exposed to vocabulary and phrasing patterns from every condition rather than a single dominant class. The domain-adapted encoder was saved and used, without further MLM training, as the initialisation for every fine-tuning experiment reported in this paper, including both baselines.

C. Model Architecture

The proposed MTL-DAPT architecture is composed of three stages, illustrated in Fig. 1. In the first stage, a single input post is tokenised and passed through the DAPT-RoBERTa encoder; the 768-dimensional [CLS] token embedding from the final transformer layer is extracted as the post's representation. In the second stage, this embedding passes through a shared encoder — a linear projection from 768 to 512 dimensions, followed by layer normalisation, a GELU activation, and dropout at probability 0.3 — producing a disorder-agnostic representation common to all three downstream tasks. This shared layer is the mechanism through which cross-condition knowledge transfer occurs: features useful for recognising depressive language remain available, unmodified in origin, to the anxiety and suicidality heads. In the third stage, three independent task

heads — one each for depression, anxiety, and suicidality — branch from the shared 512-dimensional representation. Each head consists of a linear layer to 256 dimensions, a GELU activation, dropout, and a final two-way linear output layer, producing a binary prediction per condition. The complete model contains 125,435,910 trainable parameters, with the encoder, shared layer, and all three heads updated jointly during fine-tuning.

Ground-truth binary labels for each task are derived directly from subreddit affiliation: a post drawn from r/depression yields dep = 1 and anx = sui = 0; a post from r/Anxiety yields anx = 1; a post from r/SuicideWatch yields sui = 1; and a post from r/mentalhealth yields dep = anx = sui = 0, serving as the shared negative class across all three tasks.Fig.1 shows the complete block architecture of proposed model.

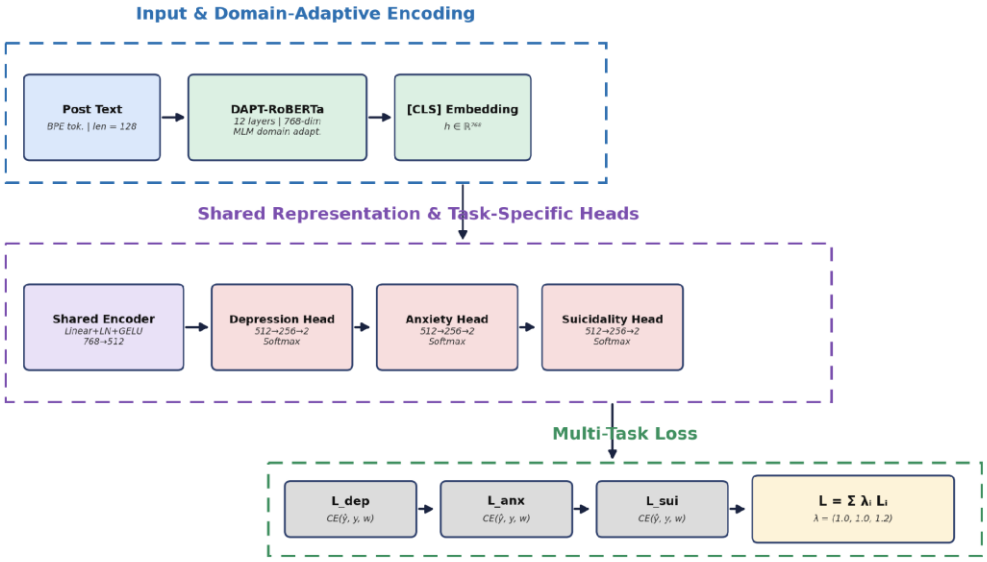

Fig. 1. MTL-DAPT architecture: DAPT-RoBERTa encoder, shared 512-dimensional encoder layer, and three independent binary task heads for depression, anxiety, and suicidality.

D. Training Strategy

The total training objective is a weighted sum of three independent binary cross-entropy losses, one per task, each additionally weighted by inverse class frequency to counteract the label imbalance documented in Section III-A:

$$L = \lambda_1 \cdot CE(\hat{y}\_dep, y\_dep, w\_dep) + \lambda_2 \cdot CE(\hat{y}\_anx, y\_anx, w\_anx) + \lambda_3 \cdot CE(\hat{y}\_sui, y\_sui, w\_sui) \quad (1)$$

where $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, and $\lambda_3 = 1.2$, with the suicidality task upweighted to reflect its comparatively greater clinical urgency [4]. Per-task class weights w_dep = [0.759, 1.465], w_anx = [0.607, 2.846], and w_sui = [0.704, 1.723] were computed on the training split via inverse class frequency (negative, positive), with anxiety receiving the largest positive-class weight owing to its smaller share of the training data. The complete model — encoder, shared layer, and all three heads — was optimised end-to-end in a single training phase using AdamW [27] with weight decay 0.01 and an initial learning rate of 2×10^{-5}, following a cosine annealing schedule across all training steps. Gradient norms were clipped to a maximum of 1.0. No separate pre-training-then-freezing phase was used for the classification stage; the encoder was fine-tuned jointly with the task heads from the first training step, consistent with standard transformer fine-tuning practice [5], [6].

IV. EXPERIMENTAL SETUP

A. Implementation Details

All experiments were implemented in PyTorch and the HuggingFace Transformers library (v4.37.2), and executed on a Kaggle notebook environment with dual NVIDIA Tesla T4 GPUs (15 GB VRAM each), using PyTorch's DataParallel module to distribute each batch across both devices. A fixed random seed (42) was used for data splitting, weight initialisation, and batch shuffling throughout. Each model — both baselines, the ablation, and the proposed MTL-DAPT model — was trained for four epochs at a per-device batch size of 128, with the maximum token length fixed at 128 sub-word tokens per post.

B. Evaluation Metrics

Model performance is reported using macro-averaged precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), computed separately for each of the three binary tasks and averaged across tasks, using the standard definitions in Eqs. (2)–(5). Confusion-matrix-derived specificity is reported alongside these to characterise negative-class performance, since aggregate accuracy alone — the metric most commonly reported in the reviewed literature [10] — cannot reveal whether a model performs uniformly across the positive and negative classes of an imbalanced task.

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall) \quad (4)$$

$$Specificity = TN / (TN + FP) \quad (5)$$

C. Baseline Models for Comparison

Two baselines and one ablation were trained and evaluated under identical data splits, batch size, epoch budget, and optimiser configuration, to isolate the specific contribution of domain adaptation from that of joint multi-task training.

Baseline 1 (BERT, single-task) is a fine-tuned bert-base-uncased [5] model with a two-layer classification head, trained on depression classification alone — representative of the majority configuration in the reviewed literature [10], with no domain adaptation and no joint training.

Baseline 2 (DAPT-RoBERTa, single-task × 3) consists of three entirely independent classifiers, each using the identical DAPT-RoBERTa encoder from Section III-B and an identical two-layer classification head, trained separately for depression, anxiety, and suicidality respectively, with no parameter sharing across the three models. This baseline uses the same domain-adapted encoder as the proposed model, differing only in whether the three tasks are trained jointly or independently, directly isolating the effect of multi-task learning.

The proposed MTL-DAPT model shares its encoder and shared 512-dimensional layer across all three tasks, as described in Section III-C, and is trained jointly under the combined loss of Eq. (1).

V. RESULTS AND DISCUSSION

A. Training Behaviour

Fig. 2 shows training and validation loss and validation macro-F1 across the four training epochs of the final MTL-DAPT model. Training loss decreases steadily and monotonically from 1.3711 at epoch 1 to 1.0502 at epoch 4, while validation loss follows a shallow U-shape, reaching its minimum at epoch 2 before rising marginally by epoch 4 — a modest divergence consistent with the onset of overfitting rather than a training instability, and one that validation macro-F1 does not reflect: F1 remains effectively flat across all four epochs (0.7902, 0.7919, 0.7923, 0.7936), indicating that the small rise in validation loss reflects growing prediction confidence on already-correct examples rather than a deterioration in classification quality. Fig2 shows the training and validation loss.

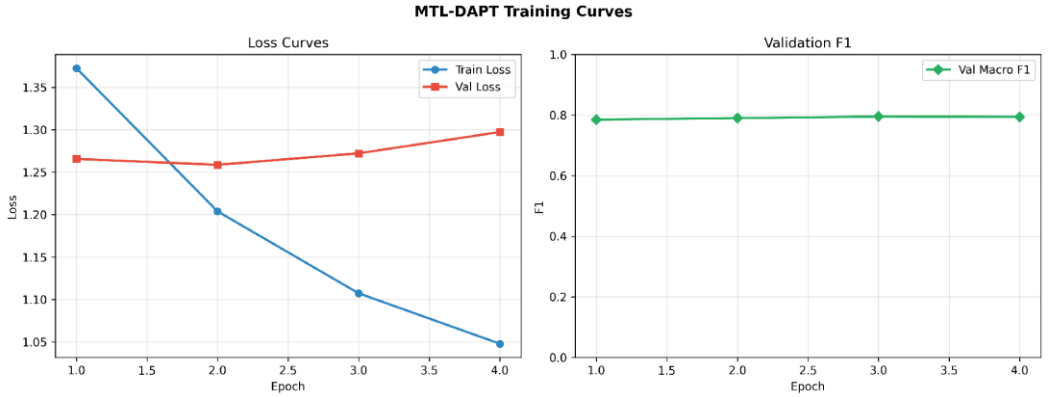

Fig. 2. Training and validation loss (left) and validation macro-F1 (right) across four epochs for the final MTL-DAPT model.

On the held-out test set of 22,500 posts, the proposed MTL-DAPT model achieves a macro-averaged F1 of 0.7970 and a macro-averaged AUC-ROC of 0.8966 across the three tasks. Table I reports the complete per-task breakdown.

B. Overall and Per-Task Test Performance

TABLE I. Per-Task Test Set Performance for MTL-DAPT (N = 22,500)

Task	Precision	Recall	F1-Score	AUC-ROC	Specificity	Support (Pos.)
Depression	0.7253	0.7488	0.7261	0.8284	0.7121	7,679
Anxiety	0.8513	0.9016	0.8732	0.9627	0.9317	3,953
Suicidality	0.7802	0.8164	0.7918	0.8988	0.8158	6,528
Macro Avg.	**0.7856**	**0.8223**	**0.7970**	**0.8966**	**0.8199**	**22,500**

Anxiety is the most reliably detected condition (F1 = 0.8732, AUC = 0.9627), followed by suicidality (F1 = 0.7918, AUC = 0.8988) and depression (F1 = 0.7261, AUC = 0.8284). This ordering is consistent with the relative lexical distinctiveness of the three subreddits: anxiety-related posts in this corpus cluster around a comparatively narrow, symptom-specific vocabulary (panic, racing heart, worry), while depression-related language overlaps considerably with the general mentalhealth control class, plausibly explaining depression's higher false-positive rate and correspondingly lower specificity (0.7121) relative to anxiety (0.9317).

C. Confusion Matrix Analysis

Fig. 3 presents the three per-task confusion matrices on the test set. For depression, 4,267 of 14,821 true-negative posts (28.8%) are misclassified as depressive, the largest false-positive rate of the three tasks, consistent with the specificity reported in Table I. For suicidality, 2,942 of 15,972 true negatives (18.4%) are similarly misclassified, while anxiety shows the lowest false-positive rate at 1,267 of 18,547 (6.8%). False-negative rates follow a comparable pattern: 1,647 of 7,679 true depressive posts (21.4%) are missed, against 508 of 3,953 (12.9%) for anxiety and 1,195 of 6,528 (18.3%) for suicidality. Taken together, these figures indicate

that the model's principal source of error is not concentrated in a single failure mode but reflects the genuine linguistic overlap between depressive and general mental-health-adjacent language documented in prior work [10].

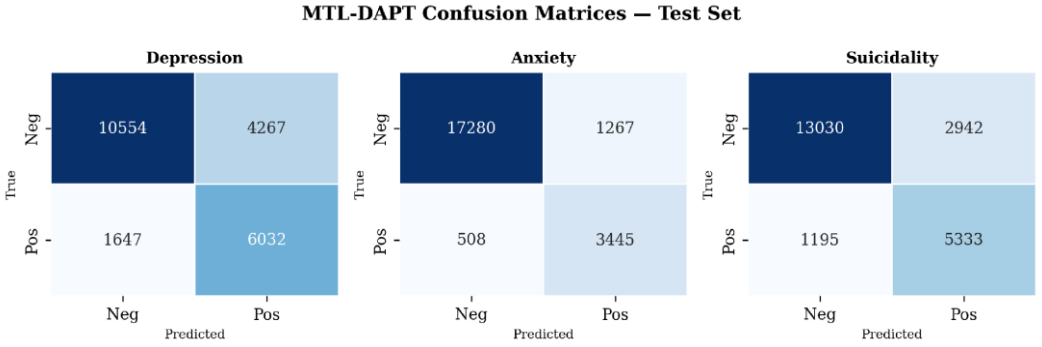

Fig. 3. Confusion matrices for depression, anxiety, and suicidality on the held-out test set (N = 22,500 per task).

D. Comparison with Baseline Models

Table II compares MTL-DAPT with the two baselines described in Section IV-C, all evaluated on the identical test split. Fig. 4 visualises the same comparison for the two DAPT-RoBERTa configurations.

TABLE II. Comparison with Baseline Models (Identical Test Split, N = 22,500)

Model	Depression F1	Anxiety F1	Suicidality F1	Macro F1	Macro AUC
BERT, single-task (depression only)	0.7120	—	—	0.7120*	0.8085*
DAPT-RoBERTa, single-task × 3 (ablation)	0.7238	0.8791	0.7909	0.7979	0.8927
MTL-DAPT (proposed)	**0.7261**	**0.8732**	**0.7918**	**0.7970**	**0.8966**

** Baseline 1 was trained on depression only; its macro F1 and macro AUC values shown are the single-task depression figures, not an average across three tasks, and are included for reference rather than direct macro-level comparison.*

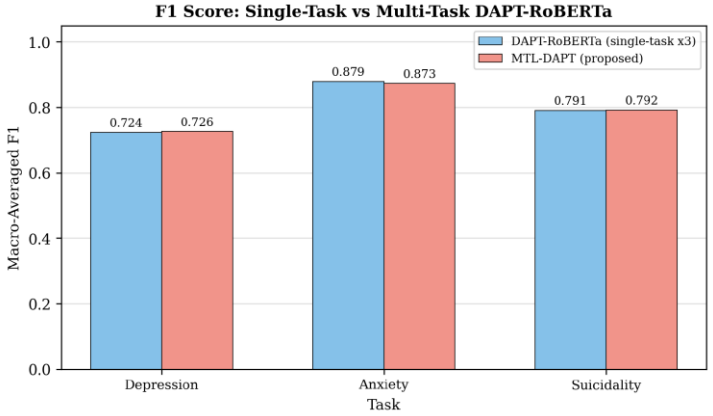

Fig. 4. F1 score comparison between single-task DAPT-RoBERTa (ablation) and the proposed MTL-DAPT model across all three tasks.

Two comparisons are of interest here. First, comparing Baseline 1 against Baseline 2 on the depression task alone isolates the contribution of domain adaptation: DAPT-RoBERTa improves F1 from 0.7120 to 0.7238 and AUC from 0.8085 to 0.8200, a gain consistent in direction and magnitude with the 1–3 F1-point improvement reported by Ji et al. [15] for MentalBERT over

general-domain BERT. Second, comparing Baseline 2 against the proposed MTL-DAPT model isolates the contribution of joint multi-task training: macro F1 changes from 0.7979 to 0.7970, a difference of −0.0009, and macro AUC improves marginally from 0.8927 to 0.8966. This difference is well within the fold-to-fold variation reported in Section V-E and is not evidence of a meaningful performance gain or loss from joint training over three independent models.

The practical significance of this result lies not in a raw performance improvement but in model efficiency: MTL-DAPT reproduces the aggregate performance of three independently trained 125-million-parameter models using a single model of the same parameter count, reducing the number of trained models required from three to one and, correspondingly, reducing inference-time cost and deployment complexity by roughly two-thirds, without any measurable degradation in per-task detection quality.

E. Three-Fold Cross-Validation

To confirm that the reported test performance reflects a stable property of the model rather than a favourable outcome of a single train–validation split, three-fold stratified cross-validation was conducted over the training partition (108,374 posts), with each fold trained for the full four-epoch budget used elsewhere in this paper. Table III reports per-fold and aggregate results.

TABLE III. Three-Fold Cross-Validation Results (MTL-DAPT)

Fold	Validation Macro F1	Validation AUC-ROC
1	0.7950	0.8935
2	0.7942	0.8916
3	0.7931	0.8913
Mean ± SD	**0.7941 ± 0.0008**	**0.8921 ± 0.0010**

A cross-validation standard deviation of 0.0008 F1 across three folds — an order of magnitude smaller than the −0.0009 difference observed between MTL-DAPT and its single-task ablation in Section V-D — indicates that the model's performance is highly stable across different partitions of the training data, and that the comparison between joint and independent training in Table II is not an artefact of a particular data split.

F. Discussion

Three findings emerge when the results above are read together. First, domain adaptation is the primary driver of the improvement this framework achieves over the standard single-task baseline: the gain from Baseline 1 to Baseline 2 (+0.0118 F1 on depression) is, in absolute terms, comparable to or larger than the difference between Baseline 2 and the proposed multi-task model. This is consistent with the broader finding that continued pre-training on in-domain text yields the majority of downstream gains in specialised NLP applications [7]. Second, joint multi-task training neither improves nor degrades per-task performance relative to independent single-task training with the same encoder — a null result in the strict sense, but one with a concrete practical implication: a single shared model can replace three independently trained ones at no measurable cost in detection quality, a property not previously reported for this combination of tasks and dataset. Third, error analysis in Section V-C indicates that depression remains the hardest of the three conditions to separate cleanly from the general mental-health-adjacent control class, a pattern that mirrors the construct-validity concerns raised in the foundational methodological review of this field [10]: subreddit affiliation is an imperfect proxy for the underlying clinical condition, and depression-related language appears to overlap more with general distress expression than either anxiety- or suicidality-related language does in this corpus.

The main limitation of this work is that ground-truth labels are inferred from subreddit affiliation rather than clinical diagnosis, a limitation shared with the vast majority of the literature reviewed in Section II [10], [25]. Moreover, all data are drawn from a single platform (Reddit) within a single three-year window. An additional limitation is the ethical consequences of this line of work — consent, privacy and downstream implications of false positives in any deployed setting — which are only partially addressed in the broader literature on inferring mental health states from social media [28] and not directly addressed by the technical contributions reported here. Cross-platform generalization and validation against clinically confirmed diagnostic labels (where such data can be ethically obtained) is left to future work.

VI. CONCLUSION AND FUTURE WORK

This paper has reported MTL-DAPT, a domain-adaptively pre-trained, multi-task learning framework for the joint detection of depression, anxiety, and suicidality from social media text, evaluated on a stratified sample of the Reddit Mental Health Dataset spanning 2020–2022. The proposed model achieves a macro-averaged F1 of 0.7970 and AUC-ROC of 0.8966 across the three tasks, with three-fold cross-validation confirming stability at 0.7941 ± 0.0008 F1. Domain adaptive pre-training accounts for the majority of the improvement over a non-adapted BERT baseline (0.7120 to 0.7238 F1 on depression), while joint multi-task training reproduces the aggregate performance of three independently trained single-task models (0.7979 macro F1) within statistical noise (−0.0009), demonstrating that a single shared architecture can replace three separate classifiers without measurable loss in per-task detection quality.

This framework should be extended in three ways in future work. First, cross-platform evaluation, training on Reddit and testing on Twitter or other social media text, would determine whether the performance reported here generalizes beyond this one platform. Second, an explicit explainability mechanism, such as attention weight visualization or integrated gradients at the token level, could be added to fill the interpretability gap highlighted in the reviewed literature [10] and provide clinically actionable insight into individual predictions, which the class-weighted loss function used in this work does not provide on its own. Third, the construct validity of the underlying ground truth would be strengthened and validation of subreddit-derived labels against clinically confirmed diagnoses, as done by Eichstaedt et al. [29] using medical record linkage, is a necessary step before any deployment in a clinical or platform-moderation context. In addition to the social media domain, transfer learning from ImageNet-scale pretraining has been shown to be similarly effective for domain-specific visual classification tasks with limited labeled data, e.g., plant disease detection [30], demonstrating domain-adaptive pretraining as a broadly transferable strategy across data modalities.

REFERENCES

[1]J. Torous, M. Keshavan, and T. Gutheil, “Promise and perils of digital psychiatry,” Asian J. Psychiatry, vol. 10, pp. 120–122, 2014, doi: 10.1016/j.ajp.2014.09.005.

[2]World Health Organization, “Mental disorders,” WHO Fact Sheet, Geneva, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3]D. A. Regier et al., “The DSM-5: Classification and criteria changes,” World Psychiatry, vol. 12, no. 2, pp. 92–98, 2013, doi: 10.1002/wps.20050.

[4]A. Abdulsalam, M. A. Abdulgabber, A. Mansour, and H. Elsayed, “Suicidal ideation detection on social media: A review of machine learning approaches,” Soc. Netw. Anal. Min., vol. 14, no. 1, p. 27, 2024, doi: 10.1007/s13278-024-01180-6.

[5]J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[6]Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.

[7]S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," in Proc. ACL, 2020, pp. 8342–8360, doi: 10.18653/v1/2020.acl-main.740.

[8]R. Caruana, "Multitask learning," Mach. Learn., vol. 28, no. 1, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[9]A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health conditions with limited social media data," in Proc. EACL, 2017, pp. 152–162, doi: 10.18653/v1/E17-1015.

[10]S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," npj Digit. Med., vol. 3, no. 1, p. 43, 2020, doi: 10.1038/s41746-020-0233-7.

[11]A. Qasim, M. Z. Asghar, and A. S. Imran, "Detection of depression severity in social media text using content-based and context-based approaches," Information, vol. 16, no. 2, p. 114, 2025, doi: 10.3390/info16020114.

[12]Q. Wan et al., "Analyzing depression in college students using NLP and transformer models," Front. Psychiatry, vol. 16, 2025, doi: 10.3389/fpsyt.2025.1576483.

[13]K. Bhatt et al., "Cross platform social media analysis for mental health detection," Discov. Ment. Health, vol. 6, no. 1, p. 45, 2026, doi: 10.1007/s44192-026-00368-w.

[14]A. Karamat et al., "A hybrid transformer architecture for multiclass mental health classification," IEEE Access, vol. 12, pp. 184279–184291, 2024, doi: 10.1109/ACCESS.2024.3512985.

[15]S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in Proc. LREC, 2022, pp. 7184–7190.

[16]I. Baydili, B. Tasci, and G. Tasci, "Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection," Behav. Sci., vol. 15, no. 3, p. 352, 2025, doi: 10.3390/bs15030352.

[17]E. Ajayi et al., "A machine learning approach for detection of mental health and cyberbullying categories from social media text," in Proc. PMLR, vol. 317, 2026.

[18]Y. Cao et al., "Machine learning approaches for mental illness detection from social media data: A systematic review," J. Behav. Data Sci., vol. 5, no. 1, pp. 24–67, 2025, doi: 10.35566/jbds/v5n1/cao.

[19]F. Kasap et al., "Ensemble of transformers for depression emotion classification," Brain Cogn., 2026, doi: 10.1007/s11571-026-10444-0.

[20]Q. Wan et al., "Analyzing depression in college students using NLP and transformer models: An ensemble approach," 2025.

[21]Y. He et al., "Mental health risk detection from social media text data: A scoping review," J. Med. Internet Res., 2026, doi: 10.2196/73945.

[22]V. Konkolics Possobom, "A systematic literature review of machine learning approaches for detecting six mental health markers," M.S. thesis, Georgia Tech., 2026.

[23]M. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," Expert Syst. Appl., vol. 133, pp. 182–197, 2019, doi: 10.1016/j.eswa.2019.04.016.

[24]J. Shen and F. Rudzicz, "Detecting anxiety on Reddit," in Proc. CLPsych Workshop, ACL, 2017, pp. 58–65.

[25]"Reddit Mental Health Dataset (RMHD)," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/entenam/reddit-mental-health-dataset

[26]N. Vindegaard and M. E. Benros, "COVID-19 pandemic and mental health consequences: Systematic review of the current evidence," Brain Behav. Immun., vol. 89, pp. 531–542, 2020, doi: 10.1016/j.bbi.2020.05.048.

[27]I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.

[28]S. Chancellor, M. Birnbaum, E. Caine, V. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in Proc. FAT*,

2019, pp. 79–88, doi: 10.1145/3287560.3287587.

[29]J. C. Eichstaedt et al., “Facebook language predicts depression in medical records,” Proc. Natl. Acad. Sci. USA, vol. 115, no. 44, pp. 11203–11208, 2018, doi: 10.1073/pnas.1802331115.

[30]S. Gupta, S. Jain, and K. C. Bandhu, “Leveraging EfficientNet-B0 for Soybean Leaf Disease Detection: A Deep Learning Perspective,” in Proc. 2025 Int. Conf. Computational, Communication and Information Technology (ICCCIT), Indore, India, 2025, pp. 637–642, doi: 10.1109/ICCCIT62592.2025.10927945.