Benchmarking Transformer-Based Language Models for Mental Health Text Classification on Social Media: A Systematic Review and Empirical Evaluation

Kamau David Muturi[1*], Duncan Nyale[1], Casper Shikali[2]

[1]*The Cooperative University of Kenya, P.O. Box 24814-00502, Karen, Nairobi, Kenya*

[2]*South Eastern Kenya University, P.O. Box 170-90200, Kitui, Kenya*

[*]*Correspondence email: muturidavid8@gmail.com*

Abstract

Mental disorders affect about 970 million people worldwide, yet few are ever screened in a clinic. Reddit posts contain people's own accounts of distress, making automated text classification an attractive, low-cost screening aid, but published results for competing transformer models use different datasets and protocols and cannot be compared. This study compared BERT, RoBERTa, MentalBERT, and DeBERTa-v3 under a single protocol to recommend a model for low-resource, multi-class screening. A systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was combined with a benchmark on Reddit posts labelled by subreddit into four classes: anxiety, depression, bipolar disorder and schizophrenia. Each model was fine-tuned with identical hyperparameters and class-weighted cross-entropy loss on a stratified sample of 20,001 posts on one NVIDIA T4 graphics processing unit (GPU), then tested on 3,001 held-out posts using macro-F1, accuracy and per-class precision and recall. Stability came from 30 bootstrap resamples of the test predictions, pairwise differences were tested with McNemar's test, and a decision matrix weighted macro-F1 (30%), domain pretraining relevance (25%), stability (15%), literature evidence (15%), efficiency (10%) and pretraining data overlap (5%). RoBERTa-base had the highest bootstrap macro-F1 (0.806), with DeBERTa-v3-base close behind (0.805). MentalBERT did not differ significantly from RoBERTa-base or BERT-base ($p > .05$). The matrix ranked MentalBERT first (8.07/10), but a sensitivity analysis showed that this depends on the judged criteria. Domain-adaptive pretraining gave no measurable advantage at this scale. Model choice should weigh accuracy, cost and domain fit together, and outputs should serve only as a triage signal.

Keywords: *mental health classification, transformer models, domain-adaptive pretraining, class imbalance, McNemar's test*

IJPP 14(9); 26 -35

1.0 Introduction

About one in eight people in the world, nearly a billion, live with a mental or substance use disorder, including some 301 million with anxiety disorders, 280 million with depression and 24 million with schizophrenia (World Health Organization [WHO], 2022). Only about a third of people with depression receive formal care. Many, though, describe their distress on social media long before they see a clinician, and that text could help flag people who need support.

On Reddit, users post in condition-specific communities (subreddits) such as r/depression or r/schizophrenia. We call this *proxy labelling*: a post's label is the community it appeared in, not a diagnosis. The label is noisy, since the poster may be a carer or have a different diagnosis, but it avoids costly manual annotation and has become the usual way to build mental health corpora (He et al., 2026). By *low-resource screening* we mean a triage tool that runs on modest hardware and needs no specialist to operate, which is where such a tool would matter most, in places such as Kenya.

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) introduced contextual representations built through self-attention, and RoBERTa (Robustly Optimized BERT Pretraining Approach; Liu et al., 2019) showed BERT was undertrained by training the same architecture longer on about 10 times more text. Neither was built for mental health language. *Domain-adaptive pretraining* (DAPT) continues a model's pretraining on unlabelled text from the target domain before fine-tuning, so words such as "manic" or "intrusive thoughts" are already well represented. Gururangan et al. (2020) found it helps most when the domain is far from the original pretraining data and labelled data are scarce. MentalBERT (Ji et al., 2022) is BERT-base with DAPT on Reddit mental health posts. DeBERTa (Decoding-enhanced BERT with disentangled attention) takes the other route. It improves the architecture by representing each token's content and position separately in attention (P. He et al., 2021), and DeBERTa-v3 adds replaced-token detection (He et al., 2023), but it has no mental health exposure. That could matter for posts whose meaning turns on word order and negation, such as "I don't think I'm manic anymore" against "I think I'm manic again". The question, then, is whether a model that has seen mental health language beats a stronger architecture that has not.

Published results cannot settle this, because they come from different datasets and protocols. A macro-F1 of 0.85 in one paper and 0.83 in another says little about which model is better. The closest multi-model studies, Bokolo and Liu (2024) on tweets(X) and Hasan et al. (2025) on Reddit, both found RoBERTa strongest, but neither included a domain-adapted model. This study adds one, tests every pairwise difference for significance and reports running cost on a single graphics processing unit (GPU) of the kind Kenyan university researchers can use. It asks:

RQ1. What does published evidence show about BERT, RoBERTa, MentalBERT and DeBERTa-v3 on mental health text classification?
RQ2. Under identical fine-tuning on the same data, do the four models' error rates differ significantly, and does MentalBERT's DAPT give it an advantage? The null hypothesis (H_0) for each pair is that both make errors at the same rate.
RQ3. Which model suits low-resource, multi-class screening once accuracy, stability, domain fit and cost are weighed together?

"About one in eight people in the world, nearly a billion, live with a mental or substance use disorder"

2.0 Materials and Methods

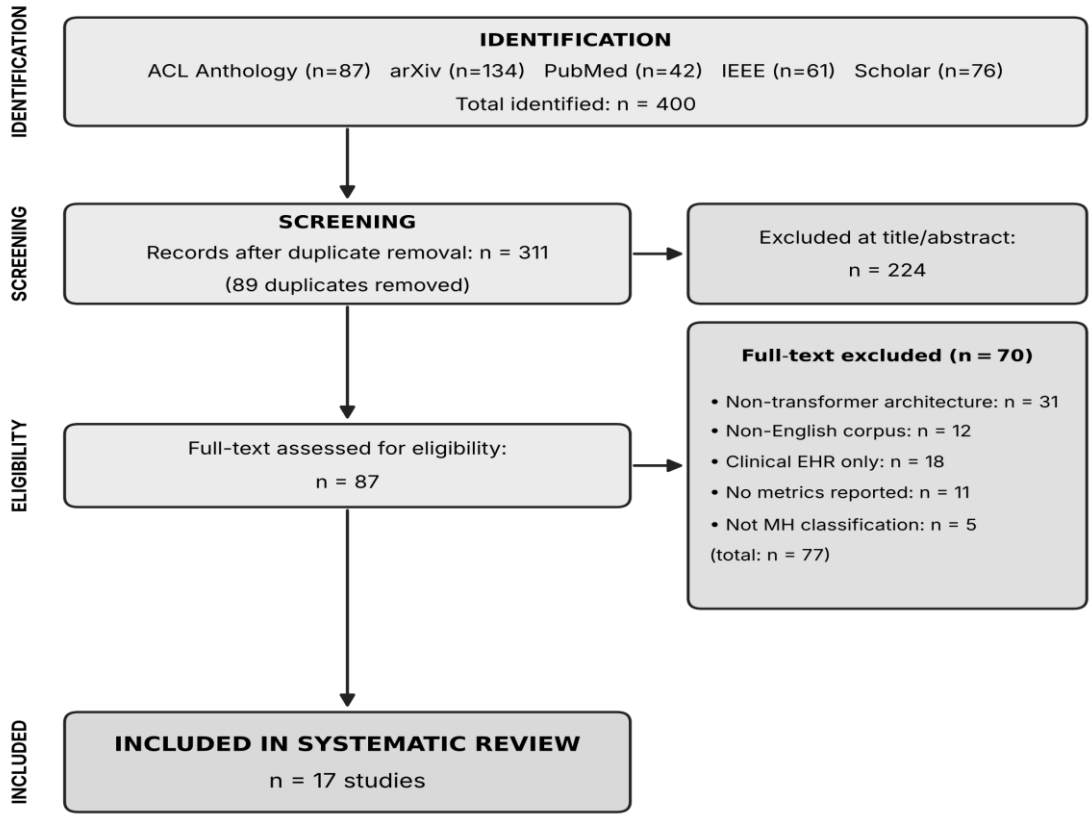
Systematic Literature Review Protocol

The review followed the PRISMA 2020 statement (Page et al., 2021). We searched ACL Anthology, arXiv, PubMed, IEEE Xplore

and Google Scholar, combining the four model names with terms such as "mental health classification", "depression detection", "Reddit" and "suicide risk detection". Studies were included if they applied a transformer to mental health classification of English social media text and reported macro-F1 or accuracy for at least one candidate model. Non-transformer, non-English and clinical-record-only studies were excluded. Of 400 records, 89 were duplicates, 224 were excluded on title and abstract, and 70 of 87 full texts were excluded, leaving 17 (Figure 1).

Figure 1

PRISMA 2020 Flow Diagram of Study Identification, Screening, Eligibility Assessment and Inclusion



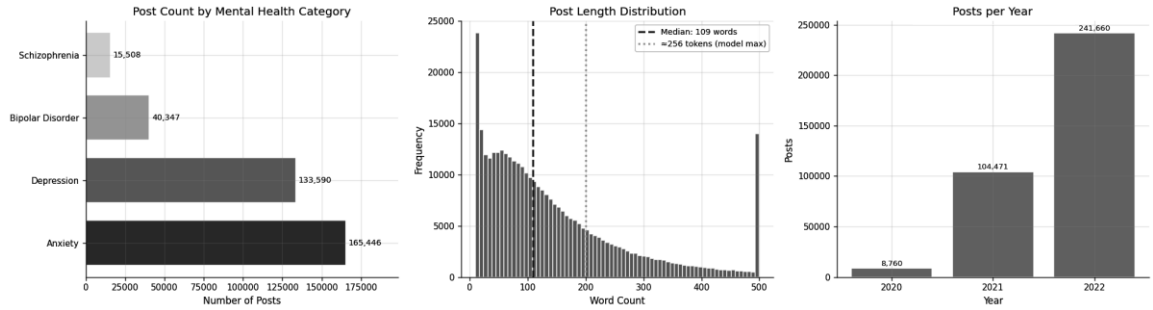
Note. Records by database: ACL Anthology (n = 87), arXiv (n = 134), PubMed (n = 42), IEEE Xplore (n = 61), Google Scholar (n = 76).

Study Dataset and Preprocessing

We used the Mental Disorders Identification (Reddit) dataset on Kaggle. Our copy held 701,787 posts from six subreddits. We excluded r/mentalillness, a general forum with no single condition, and r/BPD (borderline personality disorder), which fell outside the study's category scheme, leaving r/Anxiety, r/depression, r/bipolar and r/schizophrenia. We removed URLs, HTML entities and [deleted] or [removed] markers, lower-cased the text, joined each title to its body, and dropped posts under ten words and exact duplicates before splitting, so no text could appear in both training and test sets. The cleaned corpus held 354,891 posts (Figure 2): 46.6% Anxiety, 37.6% Depression, 11.4% bipolar disorder and 4.4% Schizophrenia. The median post had 109 words.

Figure 2

Composition of the Cleaned Corpus by Category, Post Length and Year of Posting



Note. N = 354,891 posts. Post lengths are capped at 500 words for display; the dotted line marks the approximate 256-token input limit.

Sampling and Data Splitting

Fine-tuning on the full corpus was not feasible in one Google Colab GPU session, so we drew a stratified sample of 20,001 posts and split it 70/15/15 by class with random seed 42 (Table 1). The size was set by compute, not a power calculation, but its precision can be checked. With 3,001 test posts, the standard error of an accuracy near 0.87 is about 0.006 (a 95% confidence interval of roughly ±1.2 points), so models about two macro-F1 points apart can be separated, but not models a few tenths apart. Schizophrenia, with 131 test posts, is measured far less precisely than the other classes.

Table 1

Number of Posts per Class in the Cleaned Corpus, the Stratified Sample and Each Data Split

Class	Cleaned corpus	Sample	Training	Validation	Test
Anxiety	165,446	9,324	6,526	1,399	1,399
Depression	133,590	7,529	5,270	1,129	1,130
Bipolar Disorder	40,347	2,274	1,592	341	341
Schizophrenia	15,508	874	612	131	131
Total	**354,891**	**20,001**	**14,000**	**3,000**	**3,001**

Note. Splits were stratified by class (scikit-learn, seed 42).

Candidate Models and Fine-Tuning

We fine-tuned four public Hugging Face checkpoints: bert-base-uncased (109M parameters), roberta-base (125M), mental/mental-bert-base-uncased (109M) and microsoft/deberta-v3-base (184M). Comparing MentalBERT with BERT-base isolates the effect of DAPT, while comparing RoBERTa-base and DeBERTa-v3-base with BERT-base shows the effect of better pretraining and architecture. All four used AdamW (learning rate 2×10^{-5}, weight decay 0.01, linear decay without warm-up), an effective batch size of 32, a 256-token input limit, which covers 83.4% of posts, and two epochs. Cross-entropy loss was weighted by inverse class frequency: 0.54 for Anxiety, 0.66 for Depression, 2.20 for bipolar disorder and 5.72 for Schizophrenia. Training ran on one NVIDIA T4 GPU in Google Colab with seed 42, in 16-bit mixed precision except for DeBERTa-v3-base, which was numerically unstable under 16-bit training and ran in 32-bit.

Evaluation Metrics and Statistical Testing

Macro-F1, which weights each class equally, was the primary metric. Accuracy, weighted F1 and macro precision and recall were secondary. The compute budget ruled out k-fold cross-validation, so we bootstrapped each model's 3,001 test predictions 30 times (Efron & Tibshirani, 1993). This captures sampling uncertainty in the estimate. It does not show how much the results would move if the models

were retrained. Pairwise differences were tested with McNemar's test (Dietterich, 1998), $\chi^2 = (|b - c| - 1)^2 / (b + c)$, where b and c count posts one model gets right and the other wrong, with $p < .05$ and a Holm correction across the six pairs.

Model Selection Framework

A decision matrix scored each model from 0 to 10 on macro-F1 (weight 30%), domain pretraining relevance (25%), stability (15%), literature evidence (15%), computational efficiency (10%) and pretraining data overlap with the target subreddits (5%). Macro-F1 was multiplied by 10, and stability and efficiency were mapped linearly from the best model (9.5) to the worst (6.0 and 5.0). The other three criteria were judged by the authors from the review, so we tested how much the ranking depends on them.

Ethical Considerations

The study used public, previously collected posts that contain no usernames, and it involved no contact with Reddit users. Following Benton et al. (2017) and Chancellor et al. (2019), we quote no posts, report only aggregate results and do not redistribute the data. Because a wrong label can stigmatise and a missed one can falsely reassure, the models are treated as research instruments only. Ethical approval was obtained from The Cooperative University of Kenya, and research authorization was granted by the National Commission for Science, Technology and Innovation (NACOSTI).

3.0 Results and Discussion

Systematic Review Findings

The 17 studies agree on the direction of results more than on their size. RoBERTa beat support vector machine (SVM), long short-term memory (LSTM) and BERT baselines on Reddit data (Murarka et al., 2021), RoBERTa and DeBERTa led for depression and suicide detection in tweets (Bokolo & Liu, 2024), and Hasan et al. (2025) reported F1 scores of 0.91 to 0.99 for RoBERTa across six Reddit conditions. MentalBERT outperformed general models in its own evaluation (Ji et al., 2022). For RQ1, though, these figures come from different datasets and protocols, and few studies test MentalBERT against DeBERTa-v3 directly.

Empirical Benchmarking Results

Table 2

Overall Performance, Bootstrap Stability and Computational Cost of the Four Models on the Held-Out Test Set

Model	Macro-F1	Bootstrap macro-F1, M (SD)	Accuracy	Weighted F1	Macro precision	Macro recall	Inference (ms/ batch)	GPU memory (GB)
BERT-base	0.788	0.787 (0.0076)	0.855	0.858	0.771	0.813	119.3	3.72
RoBERTa-base	0.809	0.806 (0.0081)	0.867	0.869	0.793	0.828	112.3	3.98
MentalBERT	0.797	0.795 (0.0101)	0.861	0.863	0.784	0.816	120.5	3.73
DeBERTa-v3-base	0.807	0.805 (0.0086)	0.873	0.873	0.799	0.818	652.9	10.21

Note. n = 3,001. M = mean; SD = standard deviation over 30 bootstrap resamples. Inference time is per batch of 32 on an NVIDIA T4; memory is the peak during training and inference. DeBERTa-v3-base ran in 32-bit precision, the others in 16-bit.

RoBERTa-base had the highest macro-F1 (0.809; bootstrap mean 0.806), with DeBERTa-v3-base less than one bootstrap standard deviation behind (0.807; 0.805), then MentalBERT (0.797) and BERT-base (0.788) (Table 2). MentalBERT scored about one point above BERT-base, the model it was built from. Per-class F1 tracked class size (Table 3): every model did best on Anxiety (0.897 to 0.912) and worst on Schizophrenia (0.626 to 0.686).

Table 3

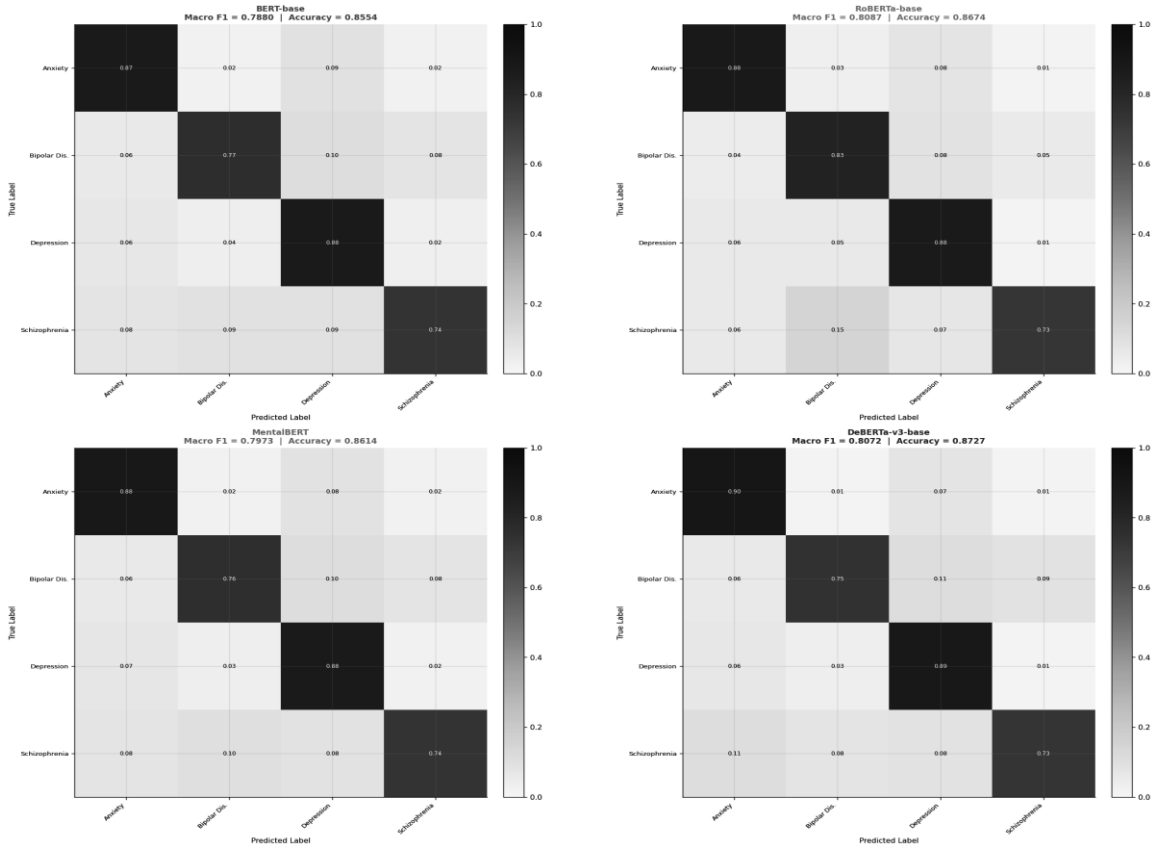
Per-Class F1 Scores of the Four Models on the Held-Out Test Set

Class	Test posts (n)	BERT-base	RoBERTa-base	MentalBERT	DeBERTa-v3-base
Anxiety	1,399	0.897	0.907	0.901	0.912
Depression	1,130	0.864	0.874	0.868	0.881
Bipolar Disorder	341	0.765	0.767	0.769	0.769
Schizophrenia	131	0.626	0.686	0.651	0.667

Note. Classes are listed from largest to smallest.

Figure 3

Row-Normalised Confusion Matrices of the Four Models on the Held-Out Test Set



Note. Each row sums to 1; diagonal cells are the share of each class classified correctly. n = 3,001.

Figure 3 shows where the errors went. RoBERTa-base and MentalBERT most often mistook Schizophrenia for Bipolar Disorder (15% and 10% of Schizophrenia posts), which fits the vocabulary the two conditions share, such as medication, hospital stays and psychotic or manic episodes. BERT-base split these errors between bipolar disorder and Depression, and DeBERTa-v3-base sent most to Anxiety (11%), so the models differ in the kind of errors they make as well as how many.

Statistical Significance

Table 4

Pairwise McNemar Tests with Full Contingency Counts for the Four Models on the Held-Out Test Set

Model A	Model B	Both correct	A correct, B wrong (b)	B correct, A wrong (c)	Both wrong	χ^2	*p*	Significant after Holm
BERT-base	RoBERTa-base	2,497	70	106	328	6.96	.008	Yes
BERT-base	MentalBERT	2,506	61	79	355	2.06	.151	No
BERT-base	DeBERTa-v3-base	2,499	68	120	314	13.84	< .001	Yes
RoBERTa-base	MentalBERT	2,510	93	75	323	1.72	.190	No
RoBERTa-base	DeBERTa-v3-base	2,520	83	99	299	1.24	.266	No
MentalBERT	DeBERTa-v3-base	2,518	67	101	315	6.48	.011	Yes

Note. n = 3,001 per comparison. χ^2 includes a continuity correction (df = 1).

Three of the six comparisons were significant, and all three survived the Holm correction (Table 4). RoBERTa-base and DeBERTa-v3-base beat BERT-base (p = .008 and p < .001), and DeBERTa-v3-base beat MentalBERT (p = .011). MentalBERT did not differ significantly from BERT-base (p = .151) or RoBERTa-base (p = .190), and RoBERTa-base did not differ from DeBERTa-v3-base (p = .266). So for RQ2, DAPT gave MentalBERT no measurable advantage. A non-significant result does not prove two models equal, though. With 140 to 188 discordant posts per pair, the test could miss small real differences.

Model Selection

Table 5

Weighted Decision Matrix Scores (0 to 10) for the Four Candidate Models

Criterion	Weight	Source	BERT-base	RoBERTa-base	MentalBERT	DeBERTa-v3-base
Macro-F1	30%	Measured	7.9	8.1	8.0	8.1
Domain pretraining relevance	25%	Judged	2.0	3.0	9.5	3.0
Result stability (bootstrap)	15%	Measured	9.5	8.9	6.0	8.1
Literature evidence	15%	Judged	9.5	9.0	6.5	5.5
Computational efficiency	10%	Measured	9.4	9.5	9.4	5.0
Pretraining data overlap	5%	Judged	1.0	1.0	9.5	1.0
Weighted total	**100%**		**6.71**	**6.87**	**8.07**	**5.77**

Note. Judged criteria were scored by the authors from the systematic review.

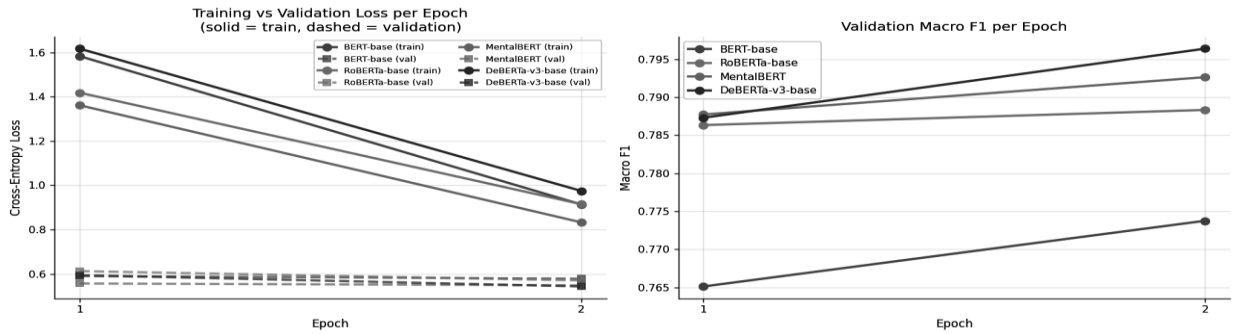
MentalBERT topped the matrix (8.07), ahead of RoBERTa-base (6.87), BERT-base (6.71) and DeBERTa-v3-base (5.77) (Table 5), but its lead comes from the two judged criteria that reward domain exposure. On the three measured criteria alone, BERT-base (8.61) and RoBERTa-base (8.57) lead and MentalBERT falls to third (7.71). BERT-base's edge there rests on a bootstrap standard deviation only 0.0005 lower, which is within sampling noise. MentalBERT stays first only while the two domain criteria carry at least about 14% of the total weight. The answer to RQ3 is therefore conditional: MentalBERT if domain fit is valued that highly, and RoBERTa-base otherwise.

Discussion

In this comparison, general model quality mattered more than domain exposure. Gururangan et al. (2020) predict the largest DAPT gains when the domain is distant and labels are scarce, and neither held here. Reddit posts are informal English, RoBERTa's pretraining data include OpenWebText, built from pages shared on Reddit (Liu et al., 2019), and 14,000 labelled posts may be enough for a general model to learn the vocabulary itself. This fits the observation that results in this field depend heavily on the dataset (Y. He et al., 2026). A learning-curve experiment would show whether MentalBERT's advantage returns with fewer labels. Disentangled attention, likewise, gave DeBERTa-v3-base no significant gain over RoBERTa-base. Validation macro-F1 was still rising at the second epoch for all four models (Figure 4), so the gaps may change with longer training.

Figure 4

Training Loss, Validation Loss and Validation Macro-F1 of the Four Models Across Two Fine-Tuning Epochs



Note. Left: training (solid) and validation (dashed) loss. Right: validation macro-F1 after each epoch.

Schizophrenia posts use distinctive language about psychosis and medication, yet the class was the hardest for every model. Volume explains this better than vocabulary: Schizophrenia was 4.4% of the training data against 46.6% for Anxiety, and per-class F1 tracked class size despite the weighted loss. Label noise adds clinical and cultural problems. Anxiety, depression and bipolar disorder often co-occur, so a post in r/depression may describe the depressive phase of bipolar disorder. And Reddit's users are mostly North American and European, while many Kenyans describe distress in bodily or spiritual terms, or in Kiswahili, Sheng or code-mixed English. A model trained on Reddit labels may therefore err systematically on other populations.

Cost matters most where screening is most needed. DeBERTa-v3-base matched RoBERTa-base statistically but took 5.8 times as long per inference batch (652.9 ms against 112.3 ms) and 2.6 times the GPU memory, partly because it had to run in 32-bit precision. BERT-base and MentalBERT cost about the same as RoBERTa-base. The study's limits are that it covered four conditions, used proxy labels, trained each model once on a 5.6% sample, and tested on a single platform. Models often lose much of their accuracy when moved to a new platform (Harrigian et al., 2020).

4.0 Conclusion

Under identical conditions, RoBERTa-base and DeBERTa-v3-base reached the highest macro-F1, about 0.81, and could not be separated statistically. MentalBERT's domain-adaptive pretraining gave no significant gain over its base model, and class size, more than each condition's language, drove per-class performance. MentalBERT ranks first only when domain fit carries at least about 14% of the decision weight; on measured performance and cost, RoBERTa-base is the stronger choice. The main contribution is a transparent way to choose among models: report accuracy, significance and running cost together, and show how far the choice rests on judgement.

5.0 Recommendations

For researchers with limited compute, a stratified split with bootstrap estimates is a workable substitute for k-fold cross-validation, provided it is reported as sampling uncertainty. Class-weighted loss should be standard for imbalanced subreddit data, though it does not remove the effect of imbalance. RoBERTa-base is a sensible default on modest hardware, MentalBERT is an option where domain fit matters, and DeBERTa-v3-base is hard to justify on a T4-class GPU. Model choices should be published with their criteria and weights.

For practice and policy, the value of these models is potential rather than proven. Before any use in Kenya, they need validation against clinician-verified labels, testing on Kiswahili and code-mixed text, and evaluation at more than one site, and their output should only prompt follow-up by a trained person, never stand as a diagnosis. Future work should use the full corpus with several seeds, run learning-curve experiments on when DAPT pays off, and test instruction-tuned large language models.

Data Availability

The dataset is publicly available on Kaggle. The fine-tuning and analysis code is available from the corresponding author on request.

References

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (94–102). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1612

Bokolo, B. G., & Liu, Q. (2024). Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. *Electronics, 13*(20), Article 3980. https://doi.org/10.3390/electronics13203980

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (79–88). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287587

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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* 1(1), 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

https://doi.org/10.1162/089976698300017197

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Harrigian, K., Aguirre, C., & Dredze, M. (2020). Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020* (3774–3788). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.337

Hasan, K., Saquer, J., & Zhang, Y. (2025). Mental multi-class classification on social media: Benchmarking transformer architectures against LSTM models. In *2025 International Conference on Machine Learning and Applications (ICMLA)* (160–167). IEEE. https://doi.org/10.1109/ICMLA66185.2025.00027

He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2111.09543

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *The Ninth International Conference on Learning Representations (ICLR 2021)*. https://doi.org/10.48550/arXiv.2006.03654

He, Y., He, Y., & Liu, D. (2026). Mental health risk detection from social media text data: A scoping review of the machine learning research landscape. *PsyCh Journal, 15*(3), Article e70100. https://doi.org/10.1002/pchj.70100

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (7184–7190). European Language Resources Association. https://aclanthology.org/2022.lrec-1.778

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021). Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis* (59–68). Association for Computational Linguistics. https://aclanthology.org/2021.louhi-1.7

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. https://www.who.int/publications/i/item/9789240049338