

Multi-Seed Evaluation of Lora-Adapted Lightweight Language Models for Malay Mental Health Text Classification

Zaaba Ahmad^{1*}, Ain Nur Balqis Hamizul¹

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia

ARTICLE INFO

Article history:

Received 27 April 2026

Revised 6 June 2026

Accepted 22 June 2026

Online first

Published 5 October 2026

Keywords:

Malay NLP

Mental health classification

Low-Rank Adaptation

Parameter-efficient fine-tuning

Reproducibility

MyDAS

DOI:

10.24191/mij.v7i2.13762

ABSTRACT

Reliable Malay mental health text classification requires accurate and efficient models evaluated without leakage. This study presents a reproducible Low-Rank Adaptation (LoRA) evaluation for four-class classification of Malay social media text into Depression, Anxiety, Stress, and No DAS Signal. An audit of the MyDAS workbook identified 2,204 usable records and five exact duplicate text keys. After deduplication, 2,199 records were divided using a stratified 70/15/15 train-validation-test protocol with no cross-split overlap. Malay-specific BERT-Tiny with LoRA ranks 16 and 32 and multilingual DistilBERT with rank 16 were evaluated over five training seeds. BERT-Tiny rank 32 achieved $88.79 \pm 0.67\%$ macro F1. Rank 16 achieved $88.59 \pm 1.30\%$ while using 87,364 trainable parameters, approximately half the 173,380 used by rank 32. DistilBERT obtained $86.05 \pm 1.06\%$ despite containing 136.2 million total parameters and requiring slower training and inference. The two BERT-Tiny means were not significantly different in an exploratory paired-seed test, whereas both exceeded DistilBERT. Malay-specific compact encoders therefore offered the stronger accuracy-efficiency trade-off, while repeated seeds revealed a practical choice between adapter economy and training stability.

1. INTRODUCTION

Depression, anxiety, and stress are frequently expressed through digital and social media, where informal posts may contain early linguistic indicators of psychological distress. Automatic analysis can help researchers understand population-level patterns and support triage-oriented tools, but a text classifier is

^{1*} Corresponding author. E-mail address: zaaba.ahmad@uitm.edu.my

not a clinical diagnosis. Its value depends on the quality of its labels, its language coverage, and the transparency of its evaluation. These requirements are especially important for mental-health applications because errors may affect vulnerable users and because social-media language is highly contextual (Raihan et al., 2024; Skianis et al., 2024).

Malay remains comparatively under-resourced in mental-health natural language processing (NLP). Expressions of distress may include colloquial spelling, abbreviations, slang, code-switching, and culturally specific emotional cues that are not represented well by English-centred resources. The MyDAS corpus addresses part of this gap by providing Malay Facebook text labelled as Depression, Anxiety, Stress, or No DAS Signal (Ahmad, Mohamed, Conway, Zakaria, et al., 2024). Earlier work with recurrent, attention-based, and transformer architectures established that Malay mental-health signals can be classified computationally, while also showing continued ambiguity among related distress categories (Ahmad et al., 2023; Ahmad, Mohamed, Conway, & Maskat, 2024; Hayati et al., 2022).

Compact language models are attractive for this setting because they reduce memory, training, and deployment requirements. Low-Rank Adaptation (LoRA) further reduces optimisation cost by freezing the pretrained encoder and learning low-rank updates in selected attention projections (Hu et al., 2022). Recent research has shown that small or efficiently adapted models can remain competitive for specialised tasks when architecture, rank, and task data are appropriately matched (Nguyen et al., 2025; Shi et al., 2026; Wang et al., 2025; Xin et al., 2024; Xu et al., 2025). However, a single strong run is insufficient evidence. Fine-tuning performance can change materially with random weight initialisation and training order, particularly for small datasets (Dodge et al., 2020; Mosbach et al., 2021).

Preliminary, unpublished experiments explored a sequence of LoRA configurations and indicated that Malay BERT-Tiny was a promising language-specific alternative to multilingual DistilBERT. The present study extends that preliminary analysis in four substantive ways. First, it audits the source workbook and removes exact duplicate texts before splitting. Second, it separates validation from final testing so that checkpoint selection does not reuse the test set. Third, it repeats each principal configuration over five training seeds and reports mean, standard deviation, and confidence intervals. Fourth, it quantifies parameter count, training time, inference latency, class-level behaviour, and the accuracy-efficiency trade-off. The resulting analysis addresses two research questions: whether increasing LoRA rank from 16 to 32 improves BERT-Tiny reliably, and whether a larger multilingual DistilBERT baseline offers sufficient benefit to justify its computational cost.

2. RELATED WORK

2.1 Malay Mental Health Text Classification

Pretrained transformer models reduce reliance on manual feature engineering by transferring contextual representations to downstream NLP tasks (Qiu et al., 2020). Malay mental-health NLP has developed from broad reviews and dialect-focused studies toward purpose-built corpora and comparative neural models. Ahmad et al. (2023) identified limited Malay datasets and language-specific resources as central barriers. Hayati et al. (2022) demonstrated depression detection on Malay dialects using GPT-3, while Ahmad, Mohamed, Conway, and Maskat (2024) compared recurrent architectures and attention mechanisms for Malay social-media mental-health detection. MyDAS subsequently provided a four-class corpus for Depression, Anxiety, Stress, and No DAS Signal, enabling controlled comparison across classical, neural, and transformer approaches (Ahmad, Mohamed, Conway, Zakaria, et al., 2024).

The wider literature reinforces the difficulty of low-resource and code-mixed mental-health classification. Because related conditions share symptoms and linguistic cues, multi-task transformer

modelling can exploit information across condition labels; however, the available evidence from Liu and Su (2024) is a preprint based on English-language social-media datasets. Skianis et al. (2024) found variable performance when depression-severity data were translated into Greek and emphasised human supervision. Kawser et al. (2025) reported that language-specific transformer representations improved depression detection for Bangla-English code-mixed text. Domain-adapted and compact mental-health models have likewise been proposed for Chinese social media, where privacy and deployability motivate smaller architectures (Xu et al., 2025). These studies support language-aware modelling but do not remove the need for leakage controls, repeated runs, and class-wise error analysis.

2.2 Parameter-Efficient Adaptation and Compact Encoders

LoRA represents a weight update as the product of two trainable low-rank matrices and applies that update to selected frozen transformer projections. The approach can retain much of the effectiveness of full fine-tuning while training only a small fraction of model parameters (Hu et al., 2022). QLoRA extends this principle through quantisation for larger models (Detrmers et al., 2023), while more recent analyses show that LoRA rank should be selected according to task complexity rather than increased automatically (Xin et al., 2024).

Compact encoders provide a complementary efficiency mechanism. BERT-Tiny reduces the depth and hidden dimension of the original BERT design, whereas multilingual DistilBERT compresses a larger multilingual encoder. A language-specific compact model may therefore trade broad multilingual coverage for more focused representations and a substantially smaller memory footprint. The Malay BERT-Tiny checkpoint used here is distributed through the Mesolitica model repository and Malaya tooling (Mesolitica, n.d.). The comparison in this study isolates the practical question of whether Malay-specific pretraining plus LoRA can outperform a considerably larger multilingual encoder on MyDAS.

2.3 Fine-Tuning Stability and Reproducibility

Transformer fine-tuning is sensitive to classifier initialisation and training-example order. Dodge et al. (2020) showed that random seeds can alter downstream results even when hyperparameters are unchanged, and Mosbach et al. (2021) connected instability to optimisation and generalisation behaviour. Reporting only the maximum score can consequently favour configurations that received a fortunate initialisation. Mean and standard deviation over matched seeds provide a more defensible comparison, although five runs still provide limited power for formal significance testing.

Data separation is equally important. The validation set guides checkpoint selection and therefore participates indirectly in model development; the final test set must remain untouched until the selected checkpoint is fixed. Exact text duplicates can also inflate performance when equivalent content appears in different splits. The journal protocol was designed around these two controls.

3. METHODOLOGY

3.1 Dataset Audit and Label Distribution

The experiments used the MyDAS experiment workbook, `dataset_final_balanced_552.csv.xlsx`. The workbook contained 2,215 rows. Eleven rows lacked a usable condition or standardised-text value, leaving 2,204 labelled records. The usable class counts were 552 Depression, 552 Anxiety, 552 Stress, and 548 No DAS Signal. This differs from the nominal description of 552 records per class and was retained transparently rather than corrected through synthetic duplication.

Texts were normalised for duplicate detection by trimming whitespace, converting to lowercase, and collapsing repeated internal whitespace. Ten rows were involved in five exact duplicate-text keys; none of those keys had conflicting labels. One instance of each duplicate key was retained, producing 2,199 unique records: 552 Depression, 552 Anxiety, 551 Stress, and 544 No DAS Signal.

Table 1. Dataset audit and leakage-controlled split

Stage	Records	Depression	Anxiety	Stress	No DAS
Workbook rows	2,215	-	-	-	-
Usable records	2,204	552	552	552	548
After exact deduplication	2,199	552	552	551	544
Training split	1,539	386	386	386	381
Validation split	330	83	83	83	81
Test split	330	83	83	83	81

Source: Authors' audit, deduplication, and stratified partitioning of the MyDAS experiment workbook

3.2 Train-Validation-Test Protocol

The deduplicated data were divided using stratified sampling with split seed 42. Seventy per cent (1,539 records) formed the training set, 15% (330) formed the validation set, and 15% (330) formed the final test set. A post-split audit found zero normalised-text overlap between training and validation, training and test, or validation and test. The split was held fixed across all models and training seeds so that differences reflected optimisation rather than different test examples.

The validation set was evaluated after every epoch. The checkpoint with the highest validation macro F1 was restored at the end of training. Only that restored checkpoint was applied to the test set. In the preliminary workflow, the same 20% partition served as both the evaluation set used for checkpoint selection and the reported final test set; the present protocol removes that source of optimistic bias.

3.3 Models and LoRA Configurations

Three configurations were selected from the exploratory experiment series. BERT-Tiny-LoRA-r16 used mesolitica/bert-tiny-standard-bahasa-cased with LoRA rank 16 and alpha 32. BERT-Tiny-LoRA-r32 increased rank to 32 and alpha to 64 while retaining the same encoder and training schedule. The mDistilBERT-LoRA-r16 baseline used distilbert-base-multilingual-cased with rank 16 and alpha 32. LoRA adapters targeted the query and value projections in BERT-Tiny and `q_lin` and `v_lin` in DistilBERT. Classification heads and LoRA weights were trainable; the remaining encoder weights were frozen.

All configurations used dropout 0.10, learning rate $3e-4$, weight decay 0.01, batch size 8, and maximum sequence length 128. BERT-Tiny models were trained for 20 epochs; DistilBERT followed its original 10-epoch schedule. Dynamic batch padding was used in the journal protocol.

Table 2. Controlled model configurations

Configuration	LoRA	Epochs	Total params	Trainable	Trainable %
BERT-Tiny-LoRA-r16	r=16, alpha=32	20	16,563,800	87,364	0.53
BERT-Tiny-LoRA-r32	r=32, alpha=64	20	16,649,816	173,380	1.04
mDistilBERT-LoRA-r16	r=16, alpha=32	10	136,216,328	888,580	0.65

Source: Authors' experimental configurations and parameter-count calculations using the Mesolitica Malay BERT-Tiny and multilingual DistilBERT checkpoints, with LoRA adaptation

3.4 Repeated-Seed Evaluation and Metrics

Each configuration was trained with seeds 42, 52, 62, 72, and 82 while the data split remained fixed. The primary measure was macro F1 because it assigns equal importance to each class. Accuracy, macro precision, and macro recall were also recorded. Class-specific precision, recall, and F1 were computed from each test prediction file, and confusion matrices were row-normalised before averaging across seeds.

Efficiency measures included total parameters, trainable parameters, wall-clock training time, and test inference latency per sample. Experiments ran on macOS 26.5.2 (Apple Silicon, arm64) using Python 3.12.13, PyTorch 2.13.0, Transformers 4.57.6, PEFT 0.20.0, Accelerate 1.14.0, scikit-learn 1.9.0, pandas 2.3.3, NumPy 2.5.2, and SciPy 1.18.0, with the Apple Metal Performance Shaders backend. Timing values characterise this environment and should be interpreted comparatively rather than as hardware-independent benchmarks. Exploratory paired t-tests compared macro F1 across matched seeds; 95% confidence intervals were calculated for paired mean differences, and Holm's procedure-controlled family-wise error across the three comparisons. Given only five seed pairs, the tests are treated as supporting evidence rather than definitive inferential claims.

3.5 Reproducibility Records

Every run saved its configuration, environment, data audit, split sizes, overlap audit, parameter counts, epoch history, final metrics, classification report, confusion matrix, predictions, training time, and inference time. The fixed split manifest and seeds 42, 52, 62, 72, and 82 allow all models to be compared on identical records. Dataset and model identifiers are recorded below so the exact experimental inputs can be verified.

The original notebook loaded an unversioned local checkpoint from Google Drive; therefore, its exact pretrained files cannot be reconstructed from the notebook alone. The present study resolves this limitation by pinning both public pretrained checkpoints.

4. RESULTS AND DISCUSSION

4.1 Historical-Protocol Reproduction

A reproduction of BERT-Tiny-LoRA-r32 using the public Mesolitica checkpoint and the preliminary 80/20 protocol produced 86.85% accuracy and 86.80% macro F1, compared with 88.44% recorded in the original notebook. The 1.64-point difference in macro F1 should not be interpreted as a contradiction because the notebook's Google Drive checkpoint was not versioned and accelerator and library versions differed. More importantly, the reproduction audit found three exact texts shared between training and the 441-record evaluation partition, and that partition was also used for checkpoint selection. These historical

scores are retained as exploratory context, while the leakage-controlled results below form the study's principal evidence.

4.2 Multi-Seed Classification Performance

Table 3 summarises the held-out test results. BERT-Tiny-LoRA-r32 achieved the highest mean macro F1 at 88.79%, with a standard deviation of 0.67 percentage points. BERT-Tiny-LoRA-r16 followed closely at 88.59% but varied more across seeds (SD = 1.30). mDistilBERT-LoRA-r16 averaged 86.05% macro F1. The same ordering appeared for accuracy, precision, and recall.

The paired mean difference between BERT-Tiny-LoRA-r32 and BERT-Tiny-LoRA-r16 was 0.20 percentage points, 95% CI [-1.27, 1.66], $t(4) = 0.37$, Holm-adjusted $p = .728$. BERT-Tiny-LoRA-r32 exceeded mDistilBERT-LoRA-r16 by 2.74 points, 95% CI [1.53, 3.94], $t(4) = 6.31$, Holm-adjusted $p = .010$. BERT-Tiny-LoRA-r16 exceeded mDistilBERT-LoRA-r16 by 2.54 points, 95% CI [0.54, 4.55], $t(4) = 3.52$, Holm-adjusted $p = .049$. These exploratory results support the advantage of Malay-specific compact pretraining over the larger multilingual baseline, but they do not support a claim that doubling LoRA rank reliably improves mean performance.

Table 3. Five-seed held-out test performance and efficiency

Configuration	Accuracy	Macro F1	Precision	Recall	Train (s)	Infer. (ms)
BERT-Tiny-LoRA-r16	88.55 ± 1.31	88.59 ± 1.30	88.74	88.59	141.5	1.71
BERT-Tiny-LoRA-r32	88.79 ± 0.68	88.79 ± 0.67	88.93	88.84	140.6	1.94
mDistilBERT-LoRA-r16	86.06 ± 1.07	86.05 ± 1.06	86.07	86.12	238.7	5.32

Source: Authors' experimental results on the held-out MyDAS test set across five training seeds

Figure 1 shows why the expanded evaluation changes the interpretation of the exploratory results. The highest individual score was 90.35% macro F1 from BERT-Tiny-LoRA-r16 at seed 72, but this configuration also produced the lowest BERT-Tiny result at seed 82. BERT-Tiny-LoRA-r32 did not attain the highest single score, yet its narrower dispersion made it the more stable configuration. A best-run comparison would therefore favour rank 16, whereas a mean-only comparison slightly favours rank 32; reporting both centre and spread reveals the actual trade-off.

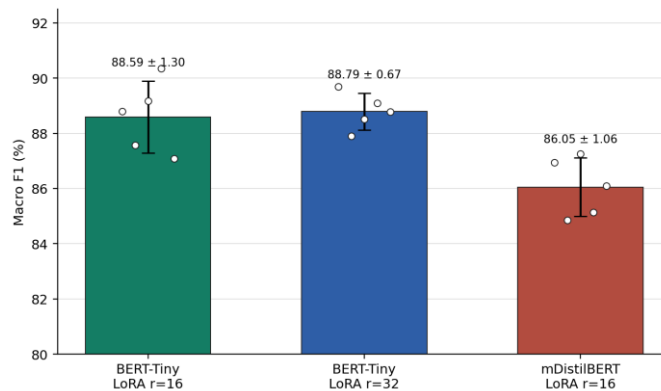


Fig. 1. Macro F1 distributions across five matched training seeds

Source: Authors' visualisation of per-seed held-out test macro F1 scores from the experiments reported in this study

4.3 Parameter and Latency Trade-Offs

The two BERT-Tiny models contained approximately 16.6 million total parameters, compared with 136.2 million for multilingual DistilBERT. BERT-Tiny-LoRA-r16 trained 87,364 parameters and BERT-Tiny-LoRA-r32 trained 173,380. Thus, doubling LoRA rank nearly doubled the trainable adapter capacity but changed mean macro F1 by only 0.20 points. Rank 16 provides the strongest parameter-efficiency result, while rank 32 provides lower seed variance.

mDistilBERT-LoRA-r16 required an average of 238.7 seconds for 10 epochs, compared with approximately 141 seconds for 20 BERT-Tiny epochs. Mean test inference latency was 5.32 ms per sample for mDistilBERT-LoRA-r16, 1.94 ms for BERT-Tiny-LoRA-r32, and 1.71 ms for BERT-Tiny-LoRA-r16. In this environment, BERT-Tiny was therefore faster despite receiving twice as many training epochs.

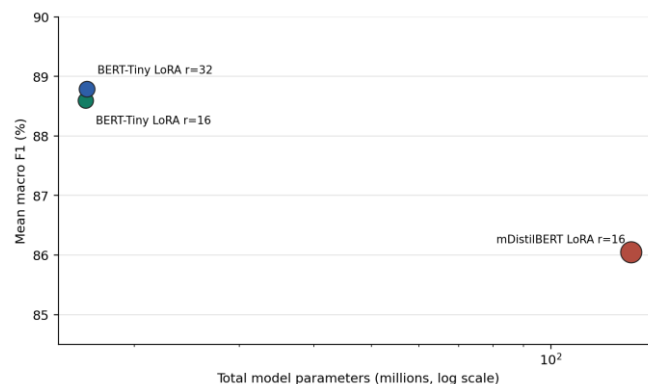


Fig. 2. Accuracy-efficiency trade-off by total model size and inference latency

Source: Authors' visualisation of mean held-out test macro F1, measured model parameter counts, and inference latency from this study

Figure 2 places the accuracy difference in operational context. Both BERT-Tiny settings occupy the favourable upper-left region: higher macro F1 with far fewer total parameters. The result aligns with recent compact mental-health modelling work showing that task and language alignment can matter more than raw model scale (Shi et al., 2026; Xu et al., 2025). It also cautions against treating parameter-efficient fine-tuning as sufficient on its own: LoRA reduces the number of updated weights, but the frozen base model still determines memory use and much of the inference cost.

4.4 Class-Level Performance and Error Patterns

No DAS Signal was the easiest class for every model. Mean class F1 reached 96.39% for BERT-Tiny-LoRA-r16, 96.09% for BERT-Tiny-LoRA-r32, and 94.05% for mDistilBERT-LoRA-r16. Distress classes were more difficult. BERT-Tiny-LoRA-r32 achieved mean F1 scores of 84.67% for Depression, 86.78% for Anxiety, and 87.62% for Stress. Its main confusion was Depression predicted as Anxiety (10.1% of Depression records), followed by Stress predicted as Anxiety (6.5%). Anxiety itself achieved 90.1% recall, indicating that the model recognised many anxiety expressions but also absorbed semantically adjacent examples.

The pattern is plausible for informal Malay mental-health text. Anxiety and stress can share descriptions of restlessness, pressure, sleep disruption, fear, and physiological arousal, while depression posts may include worry or exhaustion. The label set represents broad signals rather than mutually exclusive clinical diagnoses. Consequently, the observed confusion should guide annotation refinement and model interpretation rather than be presented as clinical differentiation.

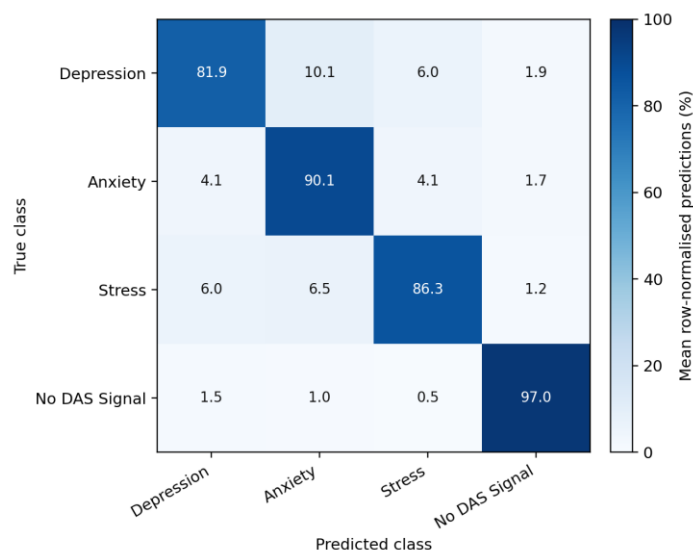


Fig. 3. Mean row-normalised confusion matrix for BERT-Tiny-LoRA-r32

Source: Authors' calculations from BERT-Tiny-LoRA-r32 held-out test predictions, with confusion matrices normalised within each true class and averaged across five training seeds

4.5 Implications for Model Selection

Model choice depends on the deployment objective. If stability across initialisations is prioritised, BERT-Tiny-LoRA-r32 is preferable because its macro F1 standard deviation was approximately half that

of BERT-Tiny-LoRA-r16. If adapter size and inference efficiency are prioritised, BERT-Tiny-LoRA-r16 is the more economical choice and its mean score is practically equivalent. mDistilBERT-LoRA-r16 offers neither a performance nor an efficiency advantage in these experiments.

The contribution therefore lies not merely in additional hyperparameter trials but in correcting evaluation boundaries, quantifying uncertainty, and connecting predictive performance to computational cost. Under this protocol, the defensible conclusion is not that a single rank is universally optimal, but that Malay-specific BERT-Tiny is robustly preferable to the multilingual baseline and that LoRA rank 16 already captures most of the attainable performance on MyDAS.

5. LIMITATIONS

The study evaluates one Malay corpus and one fixed stratified split. Five seeds improve reliability but remain a small sample for statistical inference, and external validation is needed before generalising to other platforms, dialects, age groups, or Malaysian communities. Exact deduplication does not detect paraphrases or near-duplicate posts.

MyDAS labels identify textual signals and should not be interpreted as clinical diagnoses. Social-media posts can be ambiguous, sarcastic, context-dependent, or code-mixed, and the four labels simplify potentially co-occurring conditions. Timing measurements are specific to the Apple MPS environment. Finally, the experiment does not evaluate calibration, demographic fairness, explanation quality, or robustness to adversarial spelling and slang; these remain necessary before user-facing deployment.

6. CONCLUSION

This study provides a leakage-controlled and multi-seed evaluation of LoRA-adapted lightweight models for Malay mental-health text classification. After auditing and deduplicating the MyDAS experiment workbook, three configurations were compared on a separate held-out test set. BERT-Tiny-LoRA-r32 obtained the highest mean macro F1 at 88.79%, BERT-Tiny-LoRA-r16 obtained 88.59% with approximately half as many trainable parameters, and mDistilBERT-LoRA-r16 obtained 86.05% while being substantially larger and slower.

Increasing LoRA rank from 16 to 32 reduced seed variability but did not yield a statistically reliable improvement in mean performance. The practical finding is therefore a trade-off: rank 16 is the strongest efficiency choice, while rank 32 offers greater stability. Both demonstrate that a Malay-specific compact encoder is better suited to this task than the larger multilingual baseline. Future work should add external Malay test sets, near-duplicate detection, calibration and fairness analysis, targeted annotation of Anxiety-Stress overlap, and deployment evaluation under privacy-preserving conditions.

7. ACKNOWLEDGEMENTS/FUNDING

The authors acknowledge Universiti Teknologi MARA for supporting the research environment, facilities, and academic supervision for this work. This research received no external funding.

8. CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of commercial, financial, or personal relationships that could be construed as potential conflicts of interest.

9. AUTHORS' CONTRIBUTIONS

Ain Nur Balqis Binti Hamizul: Conceptualisation, methodology, software, investigation, formal analysis, visualisation, and writing—original draft. Zaaba Bin Ahmad: Conceptualisation, supervision, validation, reproducibility review, and writing—review and editing.

10. DECLARATION OF GENERATIVE AI IN THE WRITING PROCESS

During the preparation of this work, the authors used OpenAI Codex to support language refinement, manuscript organisation, tabulation, and figure generation. The authors reviewed the experiment code, source records, calculations, citations, and manuscript content, revised the output as needed, and take full responsibility for the content of the publication.

11. DATA AVAILABILITY/SUPPLEMENTARY MATERIALS

The data analysed in this study are based on the MyDAS corpus and remain subject to its applicable access and use conditions. Experiment code, the fixed split manifest, pinned model revisions, environment details, per-seed configurations and metrics, prediction files, and generated summary tables are available from the corresponding author on reasonable request.

12. ETHICS STATEMENT

This secondary analysis used the existing MyDAS corpus and did not recruit additional participants. The original MyDAS study recruited adults, obtained explicit consent for collection and analysis of their responses and demographic data, excluded identifying information, and followed an institutionally approved ethics protocol (Ahmad, Mohamed, Conway, Zakaria, et al., 2024). The present experiments were conducted solely for language-model classification research, not clinical diagnosis. Data and experiment records were handled locally, and no identifiable or verbatim participant posts are reproduced. Any future deployment should include human oversight, privacy safeguards, transparent limitations, and referral to qualified mental-health professionals.

REFERENCES

- Ahmad, Z., Maskat, R., & Mohamed, A. (2023). Harnessing natural language processing for mental health detection in Malay text: A review. In *2023 4th International Conference on Artificial Intelligence and Data Sciences* (pp. 29–35). IEEE. <https://doi.org/10.1109/AiDAS60501.2023.10284653>
- Ahmad, Z., Mohamed, A., Conway, M., & Maskat, R. (2024). Enhancing mental health detection in Malay social media: A comparative study of RNN architectures and attention mechanisms. In *2024 5th International Conference on Artificial Intelligence and Data Sciences* (pp. 121–131). IEEE. <https://doi.org/10.1109/AiDAS63860.2024.10729940>

- Ahmad, Z., Mohamed, A., Conway, M., Zakaria, R., Ibrahim Teo, N. H., & Maskat, R. (2024). MyDAS corpus: Malay social media texts for detecting depression, anxiety, and stress on Facebook. In *2024 5th International Conference on Artificial Intelligence and Data Sciences* (pp. 111–120). IEEE. <https://doi.org/10.1109/AiDAS63860.2024.10730385>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088–10115. <https://doi.org/10.52202/075280-0441>
- Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. (2020). *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2002.06305>
- Hayati, M. F. M., Ali, M. A. M., & Rosli, A. N. M. (2022). Depression detection on Malay dialects using GPT-3. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (pp. 360–364). IEEE. <https://doi.org/10.1109/IECBES54088.2022.10079554>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. <https://openreview.net/forum?id=nZeVKeeFYf9>
- Kawser, M. S. I., Al Abrar, J., Kabir, M. B., Chowdhury, M. R., & Bahari, M. A. (2025). A hybrid transformer–sequential model for depression detection in Bangla–English code-mixed text. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)* (pp. 107–112). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.banglalp-1.8>
- Liu, J., & Su, M. (2024). *Enhancing mental health condition detection on social media through multi-task learning* [Preprint]. medRxiv. <https://doi.org/10.1101/2024.02.23.24303303>
- Mesolitica. (n.d.). *bert-tiny-standard-bahasa-cased* [Pretrained language model]. Hugging Face. <https://huggingface.co/mesolitica/bert-tiny-standard-bahasa-cased>
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. *International Conference on Learning Representations*. <https://openreview.net/forum?id=nYZ5px7Tv9>
- Nguyen, C. V., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow, J., Wu, J., Singh, A., Wang, Y., Gu, J., Ahmed, N. K., Lipka, N., Zhang, R., Chen, X., Yu, T., ... Nguyen, T. H. (2025). A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing – Natural Language Processing in the Generative AI Era* (pp. 807–821). INCOMA Ltd. <https://doi.org/10.26615/978-954-452-098-4-093>
- Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., & Huang, X. J. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), 1872–1897. <https://doi.org/10.1007/s11431-020-1647-3>
- Raihan, N., Puspo, S. S. C., Farabi, S., Bucur, A.-M., Ranasinghe, T., & Zampieri, M. (2024). MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 11196–11203). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.977/>

- Shi, J., Wang, Z., Zhou, J., Liu, C., Sun, P. Z. H., Zhao, E., & Lu, L. (2026). MentalQLM: A lightweight large language model for mental healthcare based on instruction tuning and dual LoRA modules. *IEEE Journal of Biomedical and Health Informatics*, 30(9), 8150–8161. <https://doi.org/10.1109/JBHI.2025.3594133>
- Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)* (pp. 236–241). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.mrl-1.20>
- Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2025). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6), Article 145. <https://doi.org/10.1145/3768165>
- Xin, C., Lu, Y., Lin, H., Zhou, S., Zhu, H., Wang, W., Liu, Z., Han, X., & Sun, L. (2024). Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2307–2317). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.206/>
- Xu, J., Lan, T., Valette, M., Magistry, P., & Li, L. (2025). TinyMentalLLMs enable depression detection in Chinese social media texts. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing – Natural Language Processing in the Generative AI Era* (pp. 1352–1363). INCOMA Ltd. <https://doi.org/10.26615/978-954-452-098-4-157>



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license (<http://creativecommons.org/licenses/by/4.0/>).