

MentalBERT: A Comparative Study on Addressing Mental Health Issues Using Transformer-Based Language Model

1st Piyal Saha

Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
23-51692-2@student.aiub.edu

2nd Tasmin Hasan

Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
23-52049-2@student.aiub.edu

3rd Md. Sajid Ahsan Seyam

Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
22-49010-3@student.aiub.edu

4th Zishan Ahmed Onik

Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
zishan.onik@aiub.edu

5th Kamruddin Nur

Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
kamruddin@aiub.edu

Abstract—The rapid proliferation in mental health chatbots has increased the demand for systems that accurately understand informal and emotional human conversations from real-world sources such as social media. Although clinically pre-trained models such as ClinicalBERT perform effectively in clinical environments, understanding informal language and emotional interactions in real-world mental health chatbot applications remains a challenge. This study compared the performance of MentalBERT, trained on domain-specific mental health corpora, alongside ClinicalBERT on mental health-related conversational understanding tasks. Both models were fine-tuned and evaluated under identical conditions using accuracy, precision, recall, and F1-score across the seven categories. MentalBERT consistently outperformed ClinicalBERT, achieving an F1-score of 80.74% and recall of 81.39%, representing a relative improvement of 4%. The largest gains were observed in stress and personality disorders, with notable improvements in suicidal ideation. Domain-specific pre-training on mental health data provides clear advantages over clinical domain adaptation for informal real world conversations. MentalBERT demonstrates strong potential for scalable mental health chatbot applications, particularly for early detection of depression, stress, and suicidal ideation.

Index Terms—MentalBERT, ClinicalBERT, BERT fine-tuning, Mental health chatbots, Social media text classification

and mental health-related problems. Despite these changes, hospital charts, discharge summaries, and physician-written clinical reports that prioritize objective medical observations over lived emotional experiences are still used to train the majority of clinical Natural Language Processing (NLP) models, including ClinicalBERT [1, 2]. The textual style and psychological expression of such data differ significantly from those of structured hospital documentation [3].

This study aims to demonstrate that MentalBERT significantly outperforms ClinicalBERT in practical mental health chatbot applications, where understanding mental health-related issues and the language people use in daily life is more important than explaining them [4]. This study primarily focuses on transformer-based language models for conversational mental health support systems. Specifically, we conducted this evaluation because MentalBERT, which is trained on mental health-related data, including emotionally expressive, user-generated text, consistently outperforms ClinicalBERT, which is mainly designed for electronic health records (EHR) and formal clinical documentation. Later studies discovered that models, such as MT-ClinicalBERT, improve clinical information extraction through multitask learning. However, these models are primarily designed to work with doctor-written clinical reports and notes rather than informal emotional language used by patients [5]. Although ClinicalBERT performs well on hospital-based NLP tasks, mental health chatbots engage in emotional conversations that do not rely on medical or diagnostic language [6]. This

I. INTRODUCTION

An increasing number of people express mental health-related issues outside clinical settings through digital conversations such as chatbots, online forums, and social media platforms, where people use informal language to express their feelings of fear, anxiety, depression, emotional pain,

study argues that models trained on clinical text are less suited to chatbot-based mental health applications, whereas MentalBERT is better optimized for this purpose [3].

This study makes three major contributions to the literature: First, this study systematically examines the performance of MentalBERT and ClinicalBERT in chatbot-based mental health tasks. Second, it conducts a cross-task comparison of both models using real-world mental health dialogue data. Third, it highlights why models trained on clinical text are less effective for emotional text and dialogue-based systems emphasizing the need for domain-specific pre-training.

The remainder of this paper is organized as follows. Section II includes a synthetic and comparative analysis of this study. Section III explains the dataset preprocessing and proposed the architecture. A summary of the findings is presented in Section IV. The limitations, conclusions, and future work of this study are presented in Sections V and VI, respectively.

II. LITERATURE REVIEW

Early work in NLP text classification primarily employed traditional machine learning techniques, including Random Forest, Naive Bayes, and Support Vector Machines (SVM). These approaches rely on manually engineered lexical, syntactic, and sentimental features. While effective to a limited extent, such methods often struggle to capture contextual meaning and implicit emotional expressions, particularly in informal and user-generated texts found on social media platforms. The emergence of Bidirectional Encoder Representations from Transformers (BERT) models has enabled more

context-sensitive language understanding. Devlin et al. [11] proposed BERT as a general-purpose language model trained using bidirectional self-attention, which enables improved representation of contextual relationships within the text. Several transformer variants have been developed to address the domain specific language characteristics. MentalBERT is one such model that is pre-trained on Reddit posts related to mental health. Ji et al. [4] introduced MentalBERT and evaluated its effectiveness on mental health related social media datasets.

Garg et al. [3] analyzed mental health classification tasks on social media Raihan et al. [8] investigated multi-class mental-health detection under class imbalance conditions. Early findings by Benton et al. [12] indicate that language models trained on mental health specific datasets outperform general-purpose models in mental health prediction tasks. Additional evaluations explored MentalBERT in a broader experimental setting. Garg et al. [13] assessed transformer-based architectures for wellness detection and compared MentalBERT and ClinicalBERT within ensemble and multi-model frameworks. Agarwal et al. [9] benchmarked several transformer models for depression detection and cognitive bias identification tasks using social media datasets, and reported that MentalBERT performs strongly relative to clinically trained models in informal text settings. In contrast, ClinicalBERT is pre-trained on structured clinical notes and electronic health records. Alsentzer et al. [1] introduced ClinicalBERT and demonstrated its effectiveness in several clinical NLP applications. Huang et al. [2] further examined the impact of clinical domain pre-training and showed improvements in predictive healthcare

TABLE I: Summary of Related Works

Ref.	Year	Dataset	Method	Result	Limitations
[4]	2022	Multiple (eRisk T1, CLPsych, Depression_Reddit, Dreddit, SWMH, T-SID, SAD, UMD)	MentalBERT, MentalRoBERTa	MentalBERT: 58.26–94.62% (Avg: 78.00); ClinicalBERT: 58.74–89.03% (Avg: 72.00)	English user-generated text, social media datasets, privacy and ethical concerns.
[7]	2025	Dataset is 500 annotated notes from HCPC EHR	MentalBERT, BioClinicalBERT	MentalBERT: 74%; ClinicalBERT: 72%.	Small-scale, single-site past data, Multi-label complexity, lacks live prospective validation.
[8]	2024	MentalHelp, CDS5, SWMH, SAD	flan-T5, DisorBERT, MentalBERT, RoBERTa, BERT, SuicidalBERT, ClinicalBERT, BioClinicalBERT, BioBERT, GPT-3.5-turbo	MentalBERT: 82%; ClinicalBERT: 77%.	Limited condition scope, class imbalance and single-source (Reddit) data.
[3]	2024	LoST (Reddit depression/suicidewatch)	PsychBERT, DeBERTa, DistilBERT, ClinicalBERT, MentalBERT	MentalBERT: 68.42%; ClinicalBERT: 59.14%.	The dataset is limited, the model depends on attention weights, not for clinical diagnosis.
[9]	2025	ReDepress (aakash-agarwal/ReDepress)	MentalBERT, MentalRoBERTa, ClinicalBERT, MPNet	MentalBERT F1: 0.73–0.87; ClinicalBERT F1: 0.65–0.84.	Focuses on English Reddit users, subjectivity in annotations, biases in self-reporting.
[10]	2024	MULTIWD (Reddit: 3281 posts from r/depression & r/SuicideWatch)	r/SuicideWatch	MentalBERT: 76.19%; ClinicalBERT: 73.41%.	Imbalanced data, focuses on English Reddit informal text, privacy concerns with sensitive mental health data.

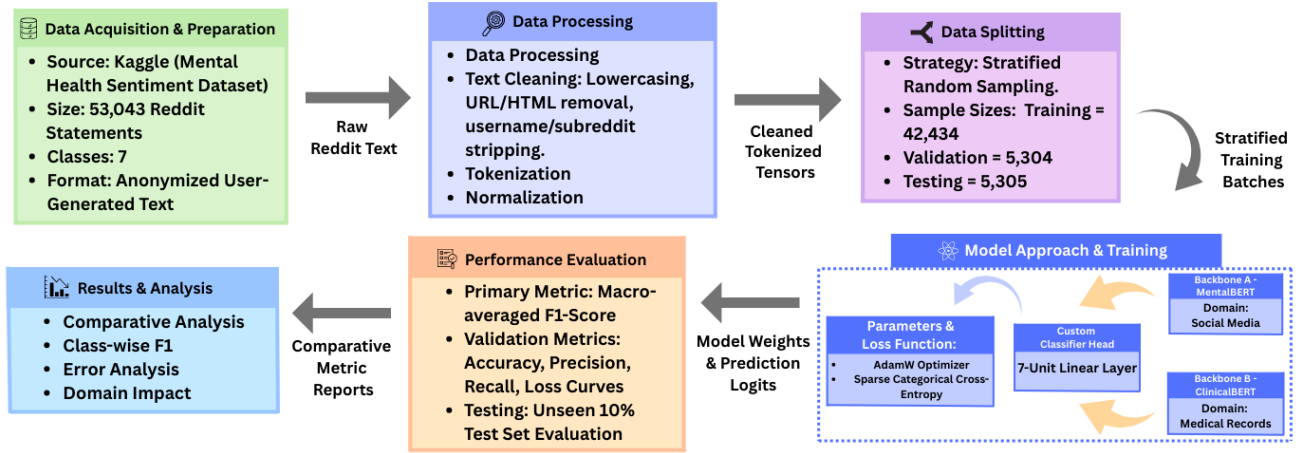


Fig. 1: Overview of the proposed methodology, starting with data preprocessing and followed by a comparative evaluation of MentalBERT and ClinicalBERT.

tasks.

However, the differences between clinical and social media texts present challenges in terms of cross-domain generalization. Ji et al. [4] compared transformer models across clinical and social media domains. Shows reduced effectiveness on informal social media content. Recent studies have explored enhancements beyond domain-specific pre-training. Kerasiotis et al. [14] examined depressive content detection and demonstrated that BERT-based models augmented with auxiliary linguistic features and synthetic data expansion outperformed MentalBERT under certain conditions. Ajayi et al. [15] evaluated transformer models fine-tuned to mental health data and reported strong performance in identifying mental health conditions and abusive language on online platforms.

Overall, existing research indicates that domain-specific transformer models outperform traditional machine-learning approaches and general-purpose BERT models in mental health-related text classification. However, only a limited number of studies have conducted direct comparisons between transformer models pre-trained on informal social media text and those pre-trained on formal clinical text for the same classification task. Consequently, the relative effectiveness of MentalBERT and ClinicalBERT when evaluated using a unified dataset remains insufficiently explored. This study addresses this research gap by directly comparing MentalBERT and ClinicalBERT for a seven-class mental health text-classification task using the same dataset. The comparison highlights the strengths and limitations of models developed with social media data versus clinical corpora and provides practical insights for real-world applications, such as conversational AI systems, automated mental health screening tools, and mental health chatbots.

III. METHODOLOGY

This study compared MentalBERT and ClinicalBERT for classifying mental health text using the “Sentiment Analysis

for Mental Health dataset” [16] from Kaggle, which contains English user-generated text.

A. Dataset Acquisition and Preparation

For this research, we employed publicly available accessible Kaggle dataset, titled “Sentiment Analysis for Mental Health” [16]. The dataset comprised 53,043 anonymized text statements, mostly collected from Reddit. Each statement was labeled into one of seven classes, encoded numerically from 0 to 6: (0) anxiety, (1) normal, (2) depression, (3) suicidal, (4) stress, (5) bipolar disorder, and (6) personality disorder.

B. Data Preprocessing and Splitting

The following preprocessing steps were applied to all text samples:

- Conversion of all text to lowercase letters.
- Removal of URLs, usernames, subreddit references, HTML tags and special characters/punctuation.
- Model-specific word-level tokenization.
- Sequence truncation and padding to a fixed length.

No data augmentation techniques were used. This decision was made consciously to avoid introducing artificial variations that could compromise the authentic expression of mental health-related language, emotional tone, and informal style of the original user-generated reddit content.

After pre-processing, the dataset was partitioned into training, validation, and test sets through stratified sampling to preserve the original class proportions.

- Training subset: 42,434 instances (approximately 80%)
- Validation subset: 5,304 instances (approximately 10%)
- Test subset: 5,305 instances (approximately 10%)

C. Model Approach and Training

A pair of domain-specific BERT-based models was further trained for seven-category mental health text classifications.

TABLE II: Dataset Statistics

Class	Train	Val	Test	Total
Anxiety	3110	389	389	3888
Bipolar Disorder	2302	288	287	2877
Depression	12323	1540	1541	15404
Normal	13081	1635	1635	16351
Personality Disorder	961	120	120	1201
Stress	2135	267	267	2669
Suicidal	8522	1065	1066	10653
Total	42434	5304	5305	53043

TABLE III: Validation Performances of ClinicalBERT vs MentalBERT

Metric	ClinicalBERT	MentalBERT
Best Epoch	5	5
Training Loss	0.7288	0.6616
Validation Loss	0.8234	0.8013
Accuracy	0.8157	0.8354
Precision	0.7751	0.8038
Recall	0.7833	0.8139
Macro F1	0.7757	0.8074

1) Backbone Network:

- **MentalBERT (mental/mentalbert-base-uncased):** This system was initially trained using mental health posts from Reddit mental health posts, allowing it to handle user-created content efficiently. It proficiently represents everyday language, emotional content, and minor intricacies that are common on online platforms.
- **ClinicalBERT (emilyalsentzer/Bio_ClinicalBERT):** This model was first trained using the healthcare documentation. Optimized for healthcare text. It reliably identifies standard clinical terms.

2) *Classifier Head:* An added fully connected output layer consisting of seven units with seven outputs was attached to every model with randomly set parameters, as reflected in the training logs. Fine-tuning proceeded for six epochs, utilizing Hugging-Face’s default configuration, except when particular parameters, such as the learning rate, batch size, and optimizer settings, were clearly adjusted.

3) *Training Strategy:* The dataset was partitioned into training, validation, and test sets using stratified sampling to preserve the class proportion, which ensured a fair representation of both majority and minority classes during training and evaluation. No additional methods, such as resampling or class weighting, were applied to address class imbalance.

D. Performance Evaluation

Model performance was measured using evaluation metrics, including loss, accuracy, precision, recall, and macro F1. The validation set was evaluated during training using an independent test set for the final evaluation. The macro-averaged F1 score was prioritized to account for imbalanced data and properly capture the minority class performance.

TABLE IV: Class-Wise F1-Score Comparison of ClinicalBERT and MentalBERT

Class	ClinicalBERT	MentalBERT
Anxiety	0.8503	0.8625
Bipolar	0.8370	0.8561
Depression	0.7593	0.7816
Normal	0.9510	0.9607
Personality Disorder	0.6417	0.6778
Stress	0.6527	0.7666
Suicidal	0.7386	0.7465

IV. RESULTS AND DISCUSSIONS

In the following section, a scientific analysis and the results of the proposed system are presented with a description of the system configurations, and the simulation results are analyzed.

A. System Configuration

Experiments were conducted on Google Colab using two NVIDIA T4 GPUs. Models were implemented in Python with PyTorch and the Hugging Face Transformers library. ClinicalBERT and MentalBERT were fine-tuned under identical settings on the same mental health text classification dataset.

B. Hyperparameters

Both models were trained with identical hyperparameters for fairness. We used the AdamW optimizer (learning rate 2×10^{-5} , weight decay 0.01, warm-up ratio 0.1) with a batch size of 16 per device. Mixed precision (fp16) and label smoothing (0.1) were applied.

C. Simulation Results

This study examined the performance of MentalBERT and ClinicalBERT within the same training setup. The best validation performance was obtained by both the models during the fifth training epoch. The training and validation performance trends are illustrated in Fig. 2, which shows a comparison of the convergence patterns of the two models.

MentalBERT and ClinicalBERT both achieved optimal validation performance by the fifth training epoch, recording a validation accuracy of 83.54% and 80.74% macro F1 score, in contrast to 81.57% and 77.57%, respectively. In addition, it reduces the training and validation losses, reflecting more effective optimization and diminished tendencies. These performance gains were preserved in the test set, where MentalBERT yielded 83.54% accuracy, 80.38% precision, 81.39% recall, and 80.74% macro F1 score. Table IV shows that MentalBERT outperformed ClinicalBERT across all classes, with the most pronounced improvements in gains in stress (+11.50%), bipolar (+2.49%), anxiety (+1.22%), and suicide (+0.81%). The training curves and test-set confusion matrices additionally validated a more rapid and stable convergence behavior and reduced number of classification errors in the underrepresented and difficult classes. Overall, the findings conclusively demonstrate that MentalBERT is the most effective model for this mental health text classification task.

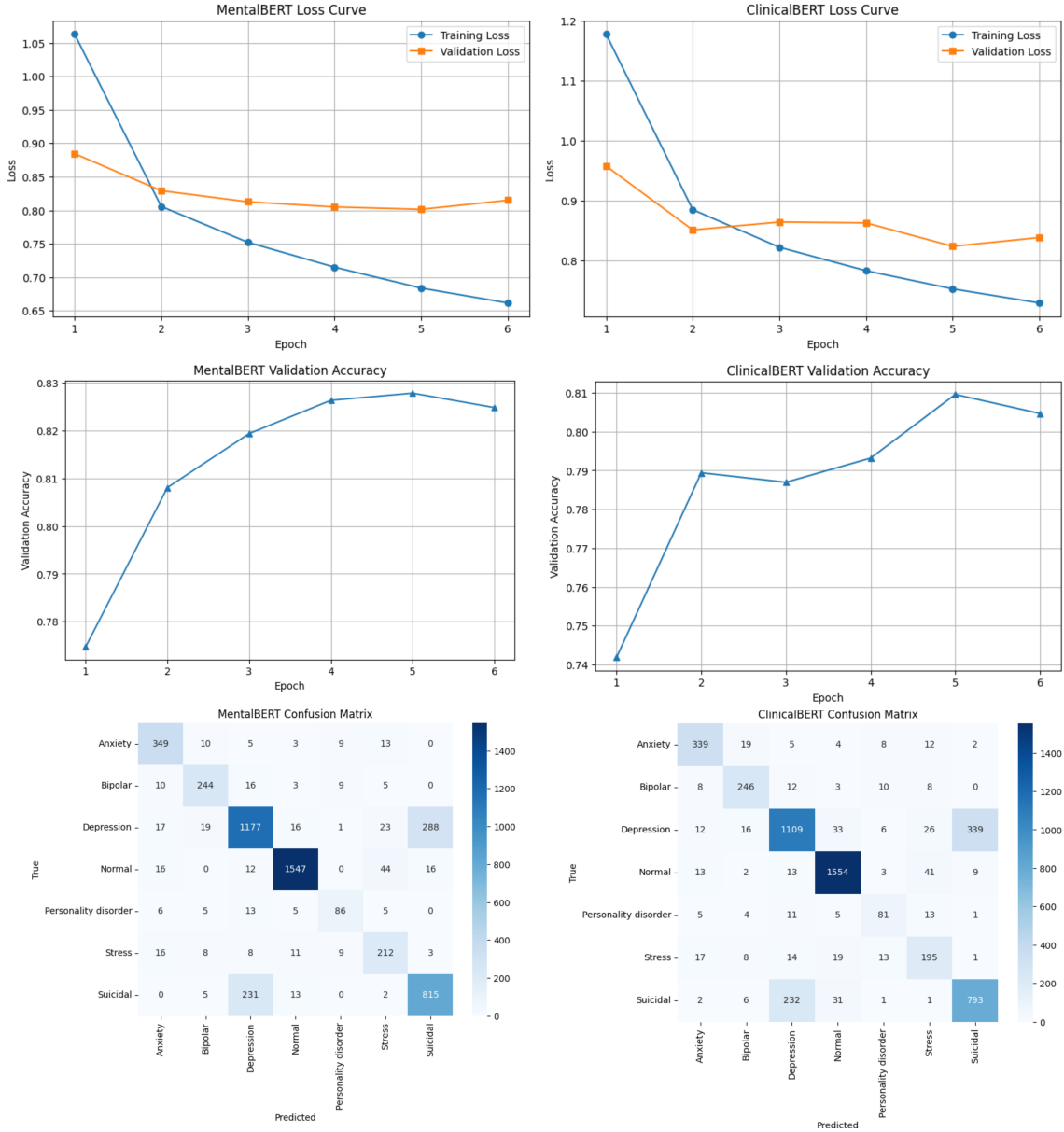


Fig. 2: Accuracy, loss curves with confusion matrix of MentalBERT and ClinicalBERT during training and validation processes.

TABLE V: Comparison of MentalBERT and ClinicalBERT Across Studies

Author	Year	Model	P (%)	R (%)	A (%)	F1 (%)	Notes
Ji et al. [4]	2022	MentalBERT	—	94.58	—	94.62	SMHD – Reddit depression detection
		ClinicalBERT	—	89.41	—	89.03	
Li et al. [7]	2024	MentalBERT	—	—	83.00	74.00	Multi-label suicide phenotyping (clinical)
		ClinicalBERT	—	—	82.00	72.00	
Garg et al. [13]	2024	MentalBERT	72.88	80.48	37.14	76.19	Multi-label wellness (Reddit posts)
		ClinicalBERT	70.86	70.70	34.25	73.41	
Agarwal et al. [9]	2025	MentalBERT	86.00	81–87	82–87	81–87	Fine-grained depression classification
		ClinicalBERT	67–83	67–84	67–85	67–84	
Our Study	2026	MentalBERT	80.38	81.39	83.54	80.74	Sentiment (Kaggle statements) – our results
		ClinicalBERT	77.52	78.33	81.58	77.58	

Input: “I feel like ending it all”

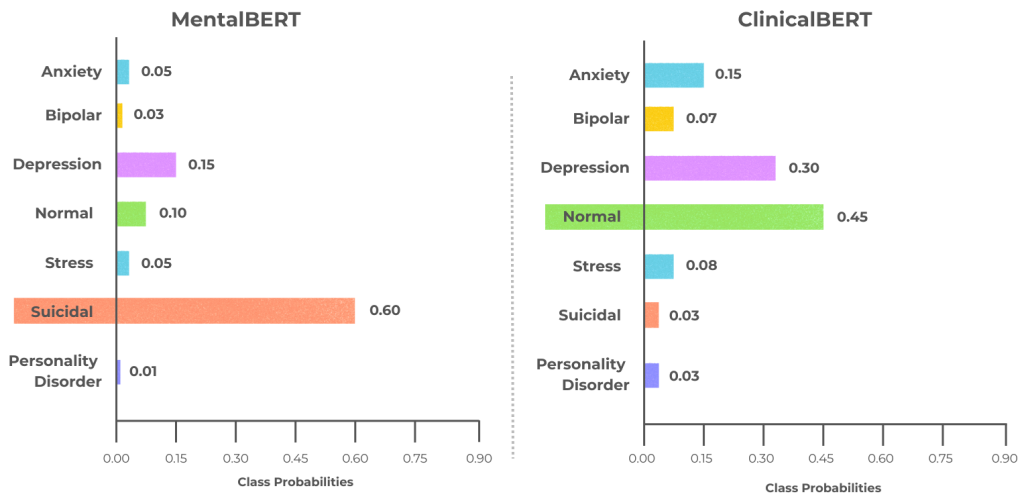


Fig. 3: Illustrative comparison of MentalBERT and ClinicalBERT predictions for a 7-class mental health sentence.

V. LIMITATIONS

A key limitation of this study is the class imbalance in the Kaggle dataset, with depression and normal classes far more frequent than stress and personality disorder. While F1 scores reduce majority class bias, overall accuracy may still favor dominant categories. The study uses a single Reddit-based dataset, limiting generalizability to other domains, such as clinical or therapeutic contexts. Standard metrics were used and do not account for clinically important priorities, such as cost-sensitive errors or higher sensitivity for high-risk categories.

VI. CONCLUSION AND FUTURE WORK

This study conducted a direct comparison of MentalBERT and ClinicalBERT for multi-class mental health text classification on a Kaggle dataset comprising anxiety, bipolar disorder, depression, normal, personality disorder, stress, and suicidal classes. MentalBERT consistently outperformed ClinicalBERT across all the major evaluation metrics, achieving 83.54% accuracy, 80.38% precision, 81.39% recall, and an 80.74% macro F1-score, whereas ClinicalBERT achieved 81.58% accuracy and a 77.58% macro F1-score. Improvements were most pronounced in the more challenging classes, including stress (+11.50%), bipolar disorder (+2.49%), anxiety (+1.22%), and suicide (+0.81%). MentalBERT also demonstrated faster convergence, lower training losses, and fewer errors in the confusion matrices, indicating stronger learning and prediction consistency. These findings underscore the advantages of pre-training on mental health-specific texts, enabling the model to better capture informal, nuanced, and emotionally rich languages. Overall, MentalBERT delivered a more reliable performance than ClinicalBERT for multi-class classification of mental health content in this dataset.

Future work will explore multilingual and diverse datasets, domain-adapted architectures, and multimodal interpretable approaches to strengthen the robustness and real-world applicability of the model for mental health support.

REFERENCES

- [1] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” in *Proceedings of the 2nd clinical natural language processing workshop*, 2019, pp. 72–78.
- [2] K. Huang, J. Altosaar, and R. Ranganath, “Clinicalbert: Modeling clinical notes and predicting hospital readmission,” *arXiv preprint arXiv:1904.05342*, 2019.
- [3] M. Garg, M. Sathvik, S. Raza, A. Chadha, and S. Sohn, “Reliability analysis of psychological concept extraction and classification in user-penned text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 422–434.
- [4] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “Mentalbert: Publicly available pretrained language models for mental healthcare,” in *proceedings of the thirteenth language resources and evaluation conference*, 2022, pp. 7184–7190.
- [5] A. Mulyar, O. Uzuner, and B. McInnes, “Mt-clinical bert: scaling clinical information extraction with multitask learning,” *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2108–2115, 2021.
- [6] A. Turchin, S. Masharsky, and M. Zitnik, “Comparison of bert implementations for natural language processing of narrative medical documents,” *Informatics in Medicine Unlocked*, vol. 36, p. 101139, 2023.
- [7] Z. Li, Y. Hu, S. Lane, S. Selek, L. Shahani, R. Machado-Vieira, J. Soares, H. Xu, H. Liu, and M. Huang, “Suicide

- phenotyping from clinical notes in safety-net psychiatric hospital using multi-label classification with pre-trained language models,” *AMIA Summits on Translational Science Proceedings*, vol. 2025, p. 260, 2025.
- [8] N. Raihan, S. S. C. Puspo, S. Farabi, A.-M. Bucur, T. Ranasinghe, and M. Zampieri, “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)*, 2024, pp. 11 196–11 203.
- [9] A. K. Agarwal, S. Bhattacharjee, M. Rastogi, J. S. Jacob, B. Banerjee, R. Gupta, and P. Bhattacharyya, “Redepress: A cognitive framework for detecting depression relapse from social media,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 34 652–34 670.
- [10] M. Sathvik and M. Garg, “Multiwd: Multiple wellness dimensions in social media posts,” *Authorea Preprints*, 2023.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [12] A. Benton, G. Coppersmith, and M. Dredze, “Ethical research protocols for social media health research,” in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 94–102.
- [13] M. Garg, X. Liu, M. Sathvik, S. Raza, and S. Sohn, “Multiwd: Multi-label wellness dimensions in social media posts,” *Journal of biomedical informatics*, vol. 150, p. 104586, 2024.
- [14] M. Kerasiotis, L. Ilias, and D. Askounis, “Depression detection in social media posts using transformer-based models and auxiliary features,” *Social Network Analysis and Mining*, vol. 14, no. 1, p. 196, 2024.
- [15] E. Ajayi, M. Kachweka, M. Deku, and E. Aiken, “A machine learning approach for detection of mental health conditions and cyberbullying from social media,” *arXiv preprint arXiv:2511.20001*, 2025.
- [16] S. Sarkar, “Sentiment analysis for mental health,” 2022, <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>.