Interpretable Depression Detection from Social Media Text Using LLM-Derived Embeddings

Samuel Kim
Department of Computer Science
Earlham College
Richmond, IN, USA
skim24@earlham.edu

Oghenemaro Imieye
Department of Computer Science
Earlham College
Richmond, IN, USA
ooimieye23@earlham.edu

Yunting Yin
Department of Computer Science
Earlham College
Richmond, IN, USA
yinyu@earlham.edu

***Abstract*—Accurate and interpretable detection of depressive language from social media is useful for early interventions, and has important implications for both clinical practice and broader public health efforts. In this paper, we investigate the performance of large language models (LLMs) and traditional machine learning classifiers across three classification tasks involving social media data: binary depression classification, depression severity classification, and differential diagnosis classification among depression, PTSD, and anxiety. We evaluate both zero-shot LLMs and machine learning models trained on traditional text embeddings and LLM-generated summary embeddings. Our experiments reveal that while zero-shot LLMs demonstrate strong generalization capabilities in binary classification, they struggle with fine-grained ordinal classifications. In contrast, classifiers trained on summary embeddings generated by LLMs demonstrate competitive, and in some cases superior, performance on the classification tasks, particularly when compared to models using traditional text embeddings. Our findings highlight the strengths of LLMs in mental health prediction, and point to promising directions for combining zero-shot capabilities with context-aware summarization techniques.**

***Index Terms*—depression detection, large language model, clinical natural language processing**

I. Introduction

Mental health disorders such as depression affect hundreds of millions of individuals worldwide, with many cases remaining undiagnosed or untreated due to social stigma, cost, or lack of access to care. As individuals increasingly express their thoughts and emotions on social media, these platforms have become a valuable source of real-time data for assessing psychological well-being. Automatic detection of depressive language from social media posts can be a promising tool for large-scale, low-cost mental health screening and intervention. Prior approaches to depression classification have typically relied on two types of features: psycholinguistic markers and text embeddings derived from pretrained language models. While psycholinguistic features such as those derived from the Linguistic Inquiry and Word Count (LIWC) lexicon offer interpretability, they are limited in expressiveness. On the other hand, traditional sentence embeddings encode rich semantic information but can be missing the specific affective cues critical for mental health prediction in this specific task.

In this work, we propose a novel prompt-based embedding approach that makes use of the reasoning capabilities of large language models (LLMs) to produce embeddings with more interpretability and semantic richness. Instead of directly embedding the raw input text, we prompt an LLM with a mental health-oriented question and feed it with users' social media post. We then extract an embedding from this LLM summary using a sentence encoder and use it as input to classifiers. This method induces reasoning by forcing the LLM to generate a semantically enriched interpretation rather than encoding surface-level syntax. It also reduces noise by allowing the LLM to filter irrelevant information. Moreover, it enhances interpretability by producing intermediate summaries that can be shown to clinicians as part of an intervention or diagnosis. We evaluate our approach on five social media-based depression datasets and find that LLM-derived summary embeddings improve predictive performance compared to models using raw text embeddings.

The remainder of this paper is organized as follows. Section II presents a review of related work, including traditional text-based approaches for mental health prediction from social media and recent applications of large language models in this domain. Section III introduces our methodology, including data preprocessing, feature extraction using text embeddings and psycholinguistic features, and the generation of LLM-based summary embeddings. Section IV reports experimental results across three classification tasks: binary depression classification, depression severity classification, and differential diagnosis among depression, anxiety, and PTSD. Finally, Section V concludes the paper with a discussion of key findings and future directions.

II. Related Works

A. Text as a Predictor of Mental Health

Textual data, whether derived from written language, transcribed speech, or online interactions, offers a powerful lens into mental health, with numerous studies demonstrating that linguistic patterns can reflect emotional states and clinical symptoms associated with mental disorders. Stress detection has been explored using a range of textual sources, including online blog and forum posts [1]–[3] and social media interactions [4]–[7]. PTSD diagnosis has also been performed using clinical patient narratives [8], [9], online surveys [10], and transcribed voicemails [11]. Sawalha et al. [12] argue that sentiment analysis of transcribed text from semi-structured virtual interviews can effectively identify individuals with PTSD using a Random Forest model with VADER sentiment scores. Building on deep learning methods, Zeberga et al. [13] present a framework for detecting depression and anxiety in social media posts using BERT and Bi-LSTM models to preserve contextual and semantic meaning, combined with a knowledge distillation approach to improve efficiency and accuracy. Mansoor et al. [14] introduces a multimodal AI model that analyzes multilingual social media data to detect early signs of mental health crises, and emphasize the need for ethical safeguards and culturally sensitive applications in real-world mental health systems. Althoff et al. [15] present a large-scale quantitative analysis of text-message-based counseling conversations using computational discourse methods. Ewbank et al. [16] develop a deep learning model to automatically categorize patient utterances during internet-enabled cognitive behavioral therapy. Bantilan et al. [17] propose an NLP model to detect suicide risk in patient messages during teletherapy, using therapist intervention patterns and expert annotations to label risk levels. These studies highlight the growing potential of natural language processing (NLP) for mental health assessment and intervention, and prove theimportance of contextual and linguistic features in the performance and real-world applicability of such models.

B. Large Language Models for Mental Health Prediction

Recent advancements in large language models have enabled their application across a wide range of domains, including the analysis and prediction of mental health conditions. Xu et al. [18] evaluates the performance of several large language models on mental health prediction tasks using online text data. Their findings suggest that while zero- and few-shot prompting yield limited results, instruction fine-tuning significantly boosts accuracy. Boggavarapu et al. [19] explore the use of LLMs enhanced with Retrieval-Augmented Generation to predict mental health-related ICD-10-CM codes from clinical notes, and find that current LLMs still struggle with accurately interpreting these complex codes. Their findings suggest the need for better integration of structured medical knowledge into these models. Malgaroli et al. [20] discuss the potential of LLMs to advance mental health care through improved diagnostics, monitoring, and treatment. They also identify challenges such as bias, accessibility, and data representation. Qian el al. [21] explore how foundation models, such as LLMs, are transforming digital mental health through personalized diagnostics, real-time monitoring, emotion recognition, and adaptive interventions using multimodal data. They propose a sociotechnical framework that integrates brain-inspired AI, and clinical oversight with ethical considerations. Hua et al. [22] review the current landscape of LLM applications in mental health care, and conclude that there is promising use cases in counseling and clinical support, but most studies lack standardized evaluation methods. Together, these studies show the growing potential of LLMs in mental health diagnosis and care. Addressing the challenges in model reliability, interpretability, and evaluation rigor will be crucial for the integration of LLMs into real-world clinical settings.

III. Methods

Our methodology is organized into four main stages: data preprocessing, psycholinguistic feature extraction, prompted LLM embedding generation, and model training and evaluation.

A. Data Preprocessing

We preprocess five publicly available social media-based mental health datasets for our experiments: MHB [23], CAMS [24], HelaDepDet [25], RMHD [26], and DepressionEmo [27]. Each dataset consists of short, user-generated text entries annotated with mental health labels, primarily related to depression. To enable evaluation of the model's ability to distinguish between depressive and non-depressive language, we also include a general-domain social media dataset, AITA [28], which contains texts unrelated to mental health, as non-depression examples in the test set.

We begin by cleaning each dataset through a series of preprocessing steps. Duplicate entries are removed, and we retain only posts with text lengths falling between the 10th and 90th percentiles to exclude outliers. For each dataset, only the columns relevant to the downstream tasks are preserved. Our experimental setup supports three classification tasks:

1) **Binary Depression Classification:** We combine all five depression-related datasets with an additional non-depression social media dataset to train models that distinguish between depressive and non-depressive content.
2) **Depression Severity Classification:** Using the HelaDepDet [25] dataset, which provides graded depression severity labels, we train models to predict levels including minimum, mild, moderate, and severe.
3) **Differential Diagnosis Classification:** We use the MHB [23] and RMHD [26] datasets, which contain multi-class annotations for depression, anxiety, and PTSD, to assess the model's ability to differentiate between related mental health conditions.

Descriptive statistics for the preprocessed datasets, including the number of posts, label categories, and average text length measured in words, are summarized in Table I. For each classification task, the compiled dataset is partitioned into a 70% training set and a 30% test set with no overlap between the two, so that traditional supervised classifiers can be trained. Notably, one of our evaluated approaches, the zero-shot LLM-based classification, does not require training data and is applied directly to the test set.

B. Text Embedding and Psycholinguistic Feature Extraction

To establish baseline performance, we evaluate classifiers trained on the combination of two types of traditional feature representations:

TABLE I: Statistics of Datasets Used Across Classification Tasks

Dataset	Used For	Size (posts)	Label Categories	Avg. Text Length (words)
MHB [1]	Binary, Differential Diagnosis	7,452	Depression, Anxiety, PTSD	253
CAMS [2]	Binary Only	4,042	Depression	179
HelaDepDet [3]	Binary, Severity Detection	33,498	Depression, Minimum, Mild, Moderate, Severe	120
RMHD [4]	Binary, Differential Diagnosis	658	Depression, Anxiety	236
DepressionEmo [5]	Binary Only	4,830	Depression	95
AITA [6]	Binary Only	24,795	Non-Depression (control)	386
Total (unique)	—	75,275	—	—

1) Text Embeddings: We extract contextualized sentence-level embeddings using the `all-mpnet-base-v2` [29] model from the SentenceTransformers library. Each social media post is passed through the pretrained model to obtain a 768-dimensional fixed-size embedding vector.

2) Psycholinguistic Features: We compute psycholinguistic features using the Linguistic Inquiry and Word Count (LIWC) lexicon. Each post is analyzed to yield normalized frequencies of relevant categories, including affective processes, cognitive processes, and pronoun usage. The resulting feature vectors are standardized using z-score normalization.

3) Classification Models: We train three traditional machine learning classifiers on the concatenation of feature sets. Logistic Regression was trained with L2 regularization (C=1.0) and a maximum of 1000 iterations. The Support Vector Machine used a linear kernel with C=1.0. The Random Forest classifier was configured with 100 decision trees and a fixed random seed to ensure reproducibility. Accuracy scores were computed on the test set to evaluate model performance.

C. Zero-Shot LLM-Based Classification

To evaluate large language models as direct zero-shot classifiers, we use the OpenAI GPT-4o API. For each social media post, we send a prompt requesting a binary or multi-class classification label, depending on the task setting. The prompt format for the binary classification task is as follows:

> You are a mental health expert. Read the following social media post and determine the user's mental health condition. Choose from the following labels: Depression, Non-depression.

For severity detection and differential diagnosis, the label options are modified accordingly. The model's textual response is parsed as the predicted label. No training or fine-tuning is applied to the LLMs, and predictions are generated directly on the held-out test set.

D. Prompted LLM Summary Embedding

In this approach, we prompt the LLM to interpret and summarize the user's mental state based on the content of a social media post. The goal is to elicit a concise, clinically oriented description that captures tone, cognitive patterns, and signals relevant to mental health assessment. The LLM-generated response is then embedded using a sentence encoder and used as input for downstream classification.

> You are a mental health expert. Read the following social media post and describe the user's mental state in one or two sentences. Focus on emotional tone, cognitive state, and any signs of mental health conditions. Avoid quoting the post verbatim.

We embed the LLM-generated response using the `all-mpnet-base-v2` [29] sentence embedding model, resulting in a 768-dimensional feature vector. This vector captures task-relevant affective semantics abstracted from the original text. We then train the same set of classifiers described in Section III-B3 on these embeddings to assess whether LLM-generated paraphrased representations improve predictive performance.

TABLE II: Binary Classification Metrics by Model and Feature Type

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (Text+LIWC)	0.89	0.94	0.89	0.91
SVM (Text+LIWC)	0.88	0.93	0.88	0.91
Random Forest (Text+LIWC)	0.91	0.96	0.90	0.93
Zero-Shot LLM-Based Classification	**0.96**	**0.97**	**0.97**	**0.97**
Logistic Regression (LLM Summary)	0.93	0.96	0.93	0.95
SVM (LLM Summary)	0.91	0.95	0.90	0.93
Random Forest (LLM Summary)	0.92	0.96	0.91	0.94

E. Evaluation Metrics

We evaluate model performance across all classification tasks using standard metrics suitable for both binary and multi-class settings. For the binary depression classification task, we report accuracy, precision, recall, and F1-score, computed on the held-out test set. For the depression severity detection and differential diagnosis classification tasks, we focus on class-wise F1-scores to get a more comprehensive understanding of model performance across categories. For all methods, including zero-shot LLM classification, we apply the same test set splits for consistency. LLM responses are parsed into discrete labels and compared to ground-truth annotations using the same metrics as the supervised models. All metrics are calculated using the `scikit-learn` library.

IV. Experiment Results

In this section, we present the evaluation results of our proposed models across three mental health classification tasks. Each task is designed to assess the model's ability

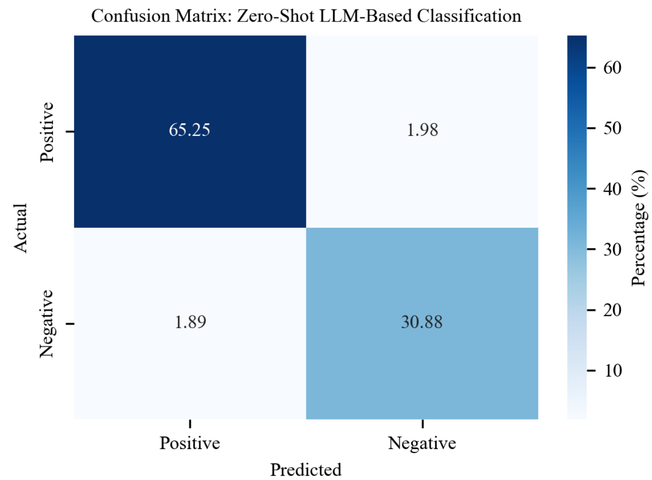
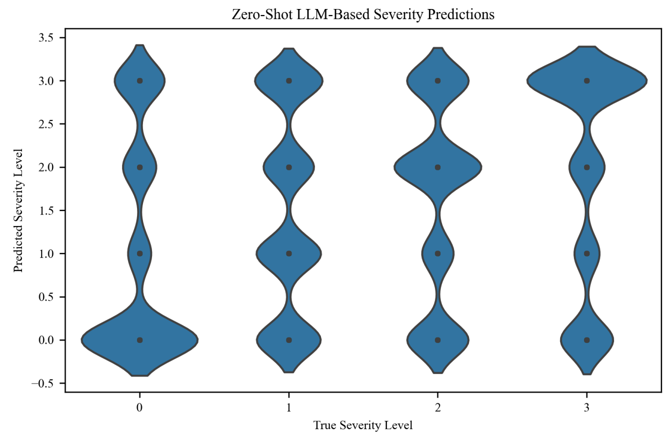
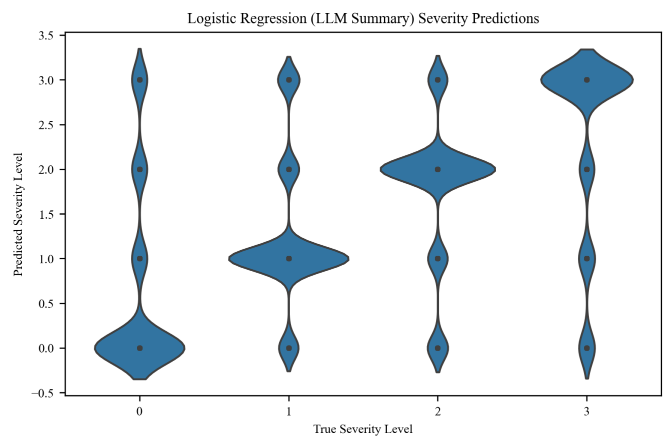


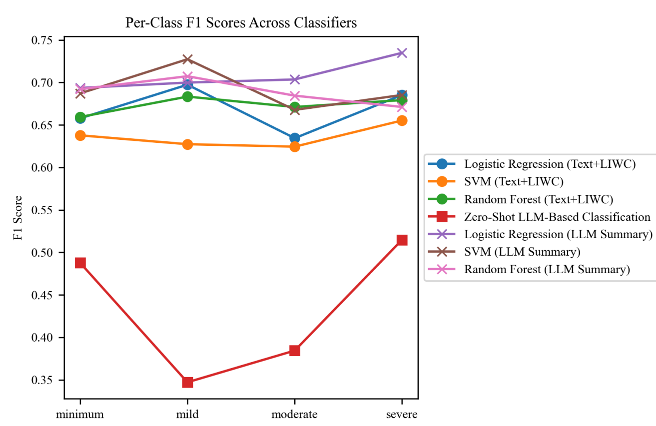
Fig. 1: Confusion matrix of the zero-shot LLM binary classifier on the test set, with percentages normalized across all predictions.



(a) Predicted depression severity levels by the zero-shot LLM-based classifier compared to true labels.



(b) Predicted depression severity levels by the Logistic Regression classifier using LLM summary embeddings compared to true labels.



(c) F1 score breakdown by depression severity level and classifier.

Fig. 2: Comparison of model performance on the depression severity classification task.

to handle different types of diagnostic complexity. We begin with binary depression classification, where the goal is to distinguish between depressive and non-depressive content using a combination of multiple datasets. Next, we examine depression severity classification, in which the models predict fine-grained severity levels using the HelaDepDet [25] dataset. Finally, we evaluate the model and embeddings' performance on differential diagnosis classification among depression, anxiety, and PTSD using the MHB [23] and RMHD [26] datasets. Performance metrics for each task are reported and analyzed in the following subsections.

A. Binary Depression Classification

For the binary classification task of determining whether a social media post is depressive or non-depressive, performance metrics are summarized in Table II. Among all models evaluated, the zero-shot LLM classifier achieved the highest overall accuracy, outperforming both traditional machine learning models using psycholinguistic and text-based embeddings, as well as models predicting based on LLM-generated summary embeddings.

Notably, the zero-shot LLM was not specifically fine-tuned for depression detection, yet demonstrated strong performance, likely due to its extensive pretraining on large-scale, diverse datasets. This proves the model's impressive generalization capabilities and supports recent findings on the excellent performance of LLMs in zero-shot settings.

The corresponding confusion matrix for the zero-shot LLM is shown in Figure 1. The model shows a low false positive rate and a low false negative rate, and is also well-balanced between precision and recall, suggesting that it does not favor minimizing one type of error at the expense of the other. However, it does show a slight tendency to classify non-depressive posts as depressive compared to other methods.

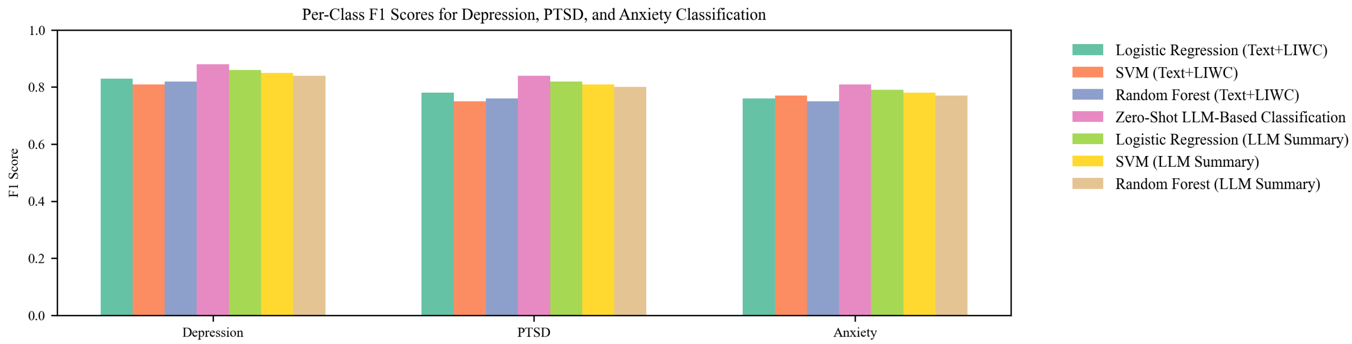


Fig. 3: Comparison of model performance on multi-class mental health diagnosis classification.

Machine learning models using LLM-generated summary embeddings performed better than those using features extracted directly from the raw social media text. This outcome is expected as the summaries provide a condensed, higher-level interpretation of each post, which likely makes implicit depressive cues more accessible.

B. Depression Severity Classification

To evaluate how well different models and embeddings can assess the severity of depression from users' social media posts, we conducted a multi-class classification task using the HelaDepDet [25] dataset. This dataset includes four ordinal labels representing increasing levels of depression severity, ranging from 0 (minimum) to 3 (severe).

We applied the same classification framework as in the previous task, comparing both zero-shot LLM-based classification and traditional machine learning models with various feature representations. Among all methods, the Logistic Regression classifier using LLM-generated summary embeddings achieved the highest accuracy of 0.58, slightly outperforming models using alternative features, and significantly outperforming the zero-shot LLM-based approach.

Figures 2a and 2b present violin plots comparing predicted and true severity levels for the best and worst-performing models. Figure 2c shows the per-class F1 scores for each classifier. We observe that the zero-shot LLM classifier struggles to infer fine-grained severity levels directly from raw text, often failing to reflect the ordinal structure of the labels. In contrast, trained machine learning models benefit from supervised learning, capturing both the semantic and ordinal relationships in the data, and are able to make accurate and consistent depression severity predictions.

C. Differential Diagnosis Classification

We evaluated model performance on the task of differential diagnosis classification using the MHB [23] and RMHD [26] datasets, which include multi-class annotations for depression, PTSD, and anxiety. This task assesses the models' ability to distinguish between related but clinically distinct mental health conditions based on social media text.

Among all evaluated methods, the zero-shot LLM-Based classifier achieved the highest overall accuracy of 65%, slightly outperforming the Logistic Regression model using LLM summary embeddings, which achieved 59% overall accuracy. The SVM classifier with text embedding and LIWC features performed the worst, with an accuracy of 47%.

Figure 3 shows the F1 score distribution across all classifiers and all three diagnostic classes. We observe that that most classifiers tend to confuse depression with anxiety, which is expected as the two mental health conditions have overlapping linguistic and emotional patterns, and social media expressions can reflect that. Moreover, PTSD is more consistently distinguished, possibly due to more specific symptom language, such as references to trauma, that sets it apart from the other two.

V. Conclusion and Future Work

This study presents a comparative evaluation of zero-shot LLMs and traditional machine learning models in depression classification tasks based on social media data. We find that zero-shot LLMs have strong performance in binary depression classification, proving their ability to generalize from pretrained knowledge. However, their performance declines in tasks requiring finer distinctions, such as severity prediction and differential diagnosis, where supervised models with LLM-generated summary embeddings show better performance.

Our evaluation results suggest that LLMs are powerful for mental health prediction tasks, and their contextual summaries are helpful to derive better features. Summary embeddings derived from LLMs capture important semantic cues and can improve traditional models to make more accurate and consistent predictions. These findings prove the potential of hybrid approaches that combine the generalization capabilities of LLMs with lightweight, interpretable classifiers trained on curated features. Further performance gains may be achieved by exploring advanced prompting strategies, applying few-shot learning , and fine-tuning LLMs on domain-specific mental health data.

Acknowledgment

This research was supported by the Lemann Student/Faculty Collaborative Research Fund at Earlham College. We gratefully acknowledge their funding and support.

References

[1] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, "Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing," Hum-Cent Intell Syst 3, 80–91, 2023.

[2] L. Zhao, J. Jia, and L. Feng, "Teenagers' stress detection based on time-sensitive micro-blog comment/response actions," In IFIP International Conference on Artificial Intelligence in Theory and Practice, pp. 26–36, 2015.

[3] B. Desmet, G. Jacobs, and V. Hoste, "Mental distress detection and triage in forum posts: the lt3 clpsych 2016 shared task system," In Proc. Third Workshop on Computational Linguistics and Clinical Psychology, pp. 148–152, 2016.

[4] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," J Big Data 9, 33, 2022.

[5] N. Li, H. Zhang, and L. Feng, "Incorporating forthcoming events and personality traits in social media based stress prediction," IEEE Trans. Affect. Comput, 2021.

[6] S. C. Guntuku, A. Buffone, L. Jaidka, J. C. Eichstaedt, and L. H. Ungar, "Understanding and measuring psychological stress using social media," In Proc. International AAAI Conference on Web and Social Media, vol. 13, pp. 214–225, 2019.

[7] M. H. Kabir, N. Samrat, A. Al Mahmud, R. Akter and M. Raihan, "Mental Stress Prediction From the Text of Social Media Using Machine Learning Techniques," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023.

[8] Q. He, B. P. Veldkamp, C. A. W. Glas, and T. de Vries, "Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining," Assessment 24, 157–172, 2017.

[9] S. Wshah S, C. Skalka C, and M. Price, "Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach," JMIR Ment Health 2019;6(7):e13946.

[10] D. Marengo, C. M. Hoeboer, B. P. Veldkamp, and M. Olff, "Text mining to improve screening for trauma-related symptoms in a global sample," Psychiatry Research, Volume 316, 2022.

[11] J. R. Oltmanns, H. A. Schwartz, C. Ruggero, Y. Son, J. Miao, M. Waszczuk, S. A. P. Clouston, E. J. Bromet, B. J. Luft, and R. Kotov, "Artificial intelligence language predictors of two-year trauma-related outcomes," Journal of psychiatric research, 143, 239–245, 2021.

[12] J. Sawalha, M. Yousefnezhad, Z. Shah, M. R. G. Brown, A. J. Greenshaw, and R. Greiner, "Detecting Presence of PTSD Using Sentiment Analysis From Text Data," Frontiers in psychiatry, 12, 811392, 2022.

[13] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T. S. Chung, "A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model," Computational intelligence and neuroscience, 2022, 7893775.

[14] M. A. Mansoor and K. H. Ansari, "Early Detection of Mental Health Crises through Artifical-Intelligence-Powered Social Media Analysis: A Prospective Observational Study," Journal of personalized medicine, 14(9), 958, 2024.

[15] T. Althoff, K. Clark, and J. Leskovec, "Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health," Transactions of the Association for Computational Linguistics, 4:463–476, 2016.

[16] M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell, "Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts," Psychotherapy research : journal of the Society for Psychotherapy Research, 31(3), 326–338, 2021.

[17] N. Bantilan, M. Malgaroli, B. Ray, and T. D. Hull, "Just in time crisis response: suicide alert system for telemedicine psychotherapy settings," Psychotherapy Research, 31(3), 289–299, 2020.

[18] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 8, 1, Article 31 (March 2024), 32 pages, 2024.

[19] L. Boggavarapu, V. Srivastava, A. M. Varanasi, Y. Lu, R. Bhaumik, "Evaluating Enhanced LLMs for Precise Mental Health Diagnosis from Clinical Notes," medRxiv, 2024, unpublished.

[20] M. Malgaroli, K. Schultebraucks, K. J. Myrick, A. A. Loch, L. Ospina-Pinillos, T. Choudhury, R. Kotov, M. De Choudhury, and J. Torous, "Large language models for the mental health community: framework for translating code to care," The Lancet. Digital health, 7(4), e282–e285, 2025.

[21] K. Qian, H. Zhang, X. Jing, B. Hu, Y. Yamamoto, and B. W. Schuller, "Foundation models for digital mental health: igniting the dawn," Medicine Plus, Volume 2, Issue 2, 2025.

[22] Y. Hua, H. Na, Z. Li, F. Liu, X. Fang, D. Clifton, and J. Torous, "A scoping review of large language models for generative tasks in mental health care," npj Digit. Med. 8, 230, 2025.

[23] S. Boinepelli, T. Raha, H. Abburi, P. Parikh, N. Chhaya, and V. Varma, "Leveraging Mental Health Forums for User-level Depression Detection on Social Media," In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5418–5427, Marseille, France. European Language Resources Association, 2022.

[24] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, and V. Mago, "CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts," In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6387–6396, Marseille, France. European Language Resources Association, 2022.

[25] Y. H. P. P. Priyadarshana, Z. Liang, and I. Piumarta, "HelaDepDet: A Novel Multi-class Classification Model for Detecting the Severity of Human Depression," In Collaboration Technologies and Social Computing: 29th International Conference, CollabTech 2023, Osaka, Japan, August 29–September 1, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, 3–18.

[26] S. Rani, K. Ahmed, and S. Subramani, "From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives," Applied Sciences, 2024; 14(4):1547.

[27] A. B. S. Rahman, H. Ta, L. Najjar, A. Azadmanesh, A. S. Gönul, "DepressionEmo: A novel dataset for multilabel classification of depression emotions," Journal of Affective Disorders, Volume 366, 2024, Pages 445-458.

[28] A. Alhassan, J. Zhang, and V. Schlegel, "'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models," In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 267–276, Marseille, France. European Language Resources Association, 2022.

[29] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, "MPNet: masked and permuted pre-training for language understanding," In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20), Curran Associates Inc., Red Hook, NY, USA, Article 1414, 16857–16867, 2020.