Identifying Severity of Depression in Forum Posts using Zero-Shot Classifier and DistilBERT Model

Zafar Sarif[1], Sannidhya Das[2], Abhishek Das[1], Md Fahin Parvej[3], Dipankar Das[3]
[1]Aliah University, Kolkata
[2]St. Xavier's College, Kolkata
[3]Jadavpur University, Kolkata
zsarifau@gmail.com

Abstract

This paper presents our approach to the RANLP 2025 Shared Task on "Identification of the Severity of Depression in Forum Posts." The objective of the task is to classify user-generated posts into one of four severity levels of depression: subthreshold, mild, moderate, or severe. A key challenge in the task was the absence of annotated training data. To address this, we employed a two-stage pipeline: first, we used zero-shot classification with *facebook/bart-large-mnli* to generate pseudo-labels for the unlabeled training set. Next, we fine-tuned a DistilBERT model on the pseudo-labeled data for multi-class classification. Our system achieved an internal accuracy of 0.92 on the pseudo-labeled test set and an accuracy of 0.289 on the official blind evaluation set. These results demonstrate the feasibility of leveraging zero-shot learning and weak supervision for mental health classification tasks, even in the absence of gold-standard annotations.

1 Introduction

In the 21st Century, mental health has become a pressing global concern, with depression identified as one of the most prevalent and disabling mental disorders. The World Health Organization (WHO) estimates that more than 280 million people globally suffer from depression, with significant impacts on quality of life, social functioning, and productivity (WHO, 2023). Depression can affect individuals of all ages and backgrounds, with symptoms often developing as early as childhood or adolescence. Left untreated, depression can lead to severe consequences, including self-harm and suicide (Friedrich, 2017). Despite growing awareness, stigma around mental illness persists, discouraging many individuals from seeking professional help in a timely manner (Clement et al., 2015).

Depression is not a binary condition but exists on a continuum of severity. It is typically categorized into subthreshold, mild, moderate, and severe levels, each requiring different clinical interventions (American Psychiatric Association, 2013). Diagnosis traditionally involves psychological evaluations conducted by trained professionals through interviews or standardized questionnaires. However, the subjective nature of symptoms, reluctance to disclose emotional distress, and limited access to mental health services—especially in low-resource settings—often delay diagnosis and treatment (Patel et al., 2018).

With the rapid growth of online platforms, individuals increasingly turn to anonymous forums for sharing mental health concerns. These platforms have become safe spaces for seeking peer and expert support. However, the sheer volume of posts makes manual monitoring by clinicians or moderators infeasible. To address this challenge, Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP) techniques offer promising solutions. By automatically analyzing linguistic patterns in user posts, ML and DL models can assess emotional states and estimate the severity of depression, enabling timely intervention and resource prioritization (Chancellor et al., 2019).

We have participated in the RANLP 2025 shared task on "Identification of the Severity of Depression in Forum Posts." Our findings of the task are presented through this paper. The objective of the task is to classify forum posts into four severity levels: subthreshold depression (*label 0*), mild (*label 1*), moderate (*label 2*), and severe depression (*label 3*). A notable challenge is the absence of annotated training data. To overcome this, we first employed zero-shot classification

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025,
pages 126–132, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-012

using the *facebook/bart-large-mnli* [1] model to generate pseudo-labels. This model is a version of the BART model (Lewis et al., 2019) that has been fine-tuned on the MultiNLI (MNLI) database. The generated pseudo-labels are then used to fine-tune a DistilBERT [2] model, optimized for multi-class classification. Finally, we have evaluated our system on the organizers' evaluation data set. Our submission achieved a decent accuracy, demonstrating the potential of semi-supervised approaches in mental health NLP tasks.

2 Related Work

The automatic detection and classification of depression severity using digital platforms has garnered substantial attention in recent years. Traditional clinical diagnosis relies heavily on self-report questionnaires such as the Hamilton Depression Rating scale (HAM-D) and the Patient Health Questionnaire (PHQ-9), which, despite their clinical validity, are limited by subjectivity and inaccessibility for real-time monitoring (Nease et al., 2002). Consequently, researchers have explored computational methods leveraging textual, vocal, and neurophysiological signals to identify not only the presence of depression but also its severity.

Multimodal approaches have proven particularly effective in assessing depression. The authors (Stepanov et al., 2018) employed speech, facial expression, and linguistic features to predict PHQ-8 scores, concluding that behavioral cues from speech were the most reliable predictors of depression severity, surpassing visual and linguistic signals in accuracy. Similarly, an article (Dibeklioglu et al., 2017) demonstrated that combining facial and head movement dynamics with vocal prosody using autoencoders achieved robust severity classification, particularly for moderate and severe depression categories.

Language-based detection methods have also shown promise. The authors (Kabir et al., 2023) proposed a clinically inspired framework (DepTweet) to label Twitter posts using DSM-5 and PHQ-9 criteria. Their annotated dataset of over 40,000 tweets allowed for the training of models such as BERT and DistilBERT to classify posts across multiple severity levels, highlighting the potential of social media text in real-world depression assessment.

Neurophysiological research has added another dimension to severity estimation. Study in a paper (Liu et al., 2023) identified neurobiological correlates of depression severity in first-episode major depressive disorder using gamma-band EEG responses. They observed that 40 Hz and 60 Hz Auditory Steady State Responses (ASSRs) significantly correlated with clinical severity, suggesting these measures as potential diagnostic biomarkers. Similarly, another work (Mahato et al., 2020) demonstrated the effectiveness of EEG-derived features like wavelet energy and asymmetry measures in both detection and severity scaling of depression, achieving high classification accuracy using Support Vector Machine (SVM) based models.

While many prior works have relied on clinically labeled data or multimodal inputs, challenges remain in applying such techniques to forum or social media text, where class labels are often unavailable. To address this, zero-shot learning has emerged as a viable solution. In the absence of labeled data, zero-shot classification models such as *facebook/bart-large-mnli* have been utilized to generate pseudo-labels, which are then used to fine-tune more efficient downstream models like DistilBERT. This two-stage approach forms the foundation of our methodology, enabling the construction of supervised models from unlabeled depression forum datasets.

Collectively, these studies underline the effectiveness of machine learning models in detecting depressive symptoms and classifying their severity across various modalities and data sources. Our work contributes to this growing field by applying zero-shot learning and transformer fine-tuning techniques in a low-resource, text-only setting aligned with real-world use cases.

3 Datasets

The dataset for the RANLP 2025 Shared Task on 'Identification of the Severity of Depression in Forum Posts' consists of two main components: an unlabeled training dataset and an evaluation dataset intended for system testing. Both datasets contain user-generated content collected from mental health-related online forums.

[1] https://huggingface.co/facebook/bart-large-mnli

[2] https://huggingface.co/distilbert/distilbert-base-uncased

	user	text	predicted_severity
4516	SteveLC777	The first step to getting help. I have been o...	subthreshold depression
4517	Dagebow	Slip sliding away. Hi all. New to the forum ...	mild depression
4518	Easy_D	Slip sliding away. So I'm writing this to giv...	subthreshold depression
4519	thadeedz	My depression situation - anyone out there wit...	severe depression
4520	ccdep71	Lost. I have been suffering depression and an...	severe depression

Figure 1: Training dataset after generating pseudo-labels by using zero-shot classification.

The training data, distributed as a tab-separated file titled *Training_data_D-severity.tsv*, comprises 4536 entries. Each entry includes three textual fields: a user identifier (*user*), a short post heading (*title*), and a longer description (*question*). To prepare the data for model training, we merged the *'title'* and '*question'* fields into a single input string '*text'*. This dataset does not include any severity labels, which required us to generate pseudo-labels through zero-shot classification (Figure 1). The data reflects typical user-generated texts from mental health forums, featuring informal language, emotional expressions, and variability in length and structure.

The evaluation dataset, provided in a CSV file titled *evaluation_textonly.csv*, contains 5189 user posts. Each entry consists of a single column (*text*), representing the full body of a forum post. Unlike the training data, the evaluation set does not include any identifiers or structured fields such as title or question, and the gold labels indicating severity levels are not released to participants. Instead, predictions submitted by participants are scored externally by the organizers against the hidden gold labels. This setup reflects a real-world application scenario, where systems are expected to classify the severity of depression in anonymous, unstructured forum posts without access to contextual or metadata cues.

4 Methodology

Our approach to the shared task involves two-major phases - (1) generating pseudo-labels for the unlabeled training data using zero-shot classification, and (2) training a supervised text classification model using a fine-tuned transformer architecture on the pseudo-labeled data (Figure 2).

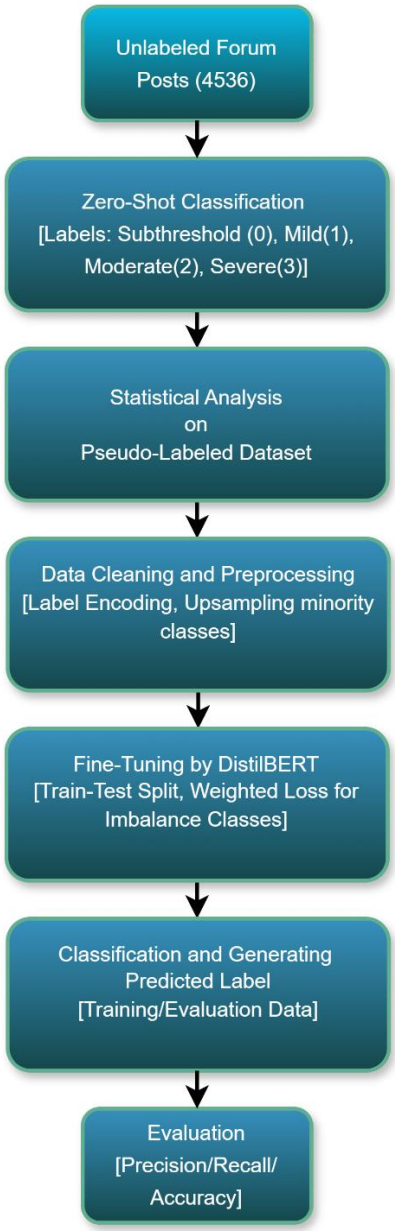

Figure 2: Pipeline of Proposed Approach.

4.1 Pseudo-label Generation

The training data was provided without severity labels. To overcome this limitation, we employed the *facebook/bart-large-mnli* model from Hugging Face Transformers for zero-shot classification (Yin et. al., 2019; Schopf et. al., 2022). For each forum

post, the title and question fields were concatenated to form a complete input. We defined four candidate labels — subthreshold depression *(label 0)*, mild depression *(label 1)*, moderate depression *(label 2)*, and severe depression *(label 3)* — and allowed the model to assign the most probable label to each post based on natural language inference. The predictions were stored and used to create a pseudo-labeled dataset.

4.2 Statistical Analysis of Data

After label generation, we have conducted some statistical analysis on data to understand the behavior and relationships among the variables.

From the label distribution graph (Figure 3), it's clearly understandable that the data has imbalance and for model fitting we need to apply some data-imbalance handling techniques. The dataset contains 63.16% severe cases, which means most of the users have a serious need for some consultation and help to deal with their problems. Severe depression is majority and subthreshold is minority class here.

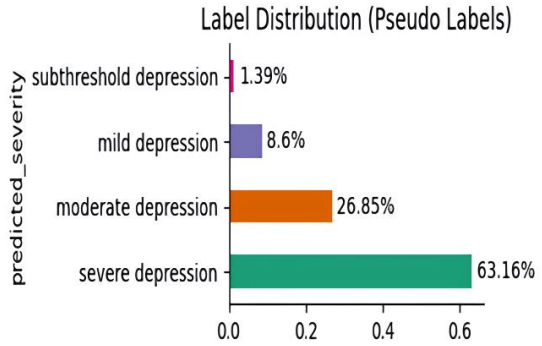

Figure 3: Label Distribution (Training Data).

A distribution of text lengths with various class shows that most posts fall between 100–400 words (Figure 4). A steep drop-off after 500 words is observed, but some posts go up to 1900+ words also (outliers). This long tail implies that a small number of users write significantly longer posts. That's why, text with ~512 tokens are considered during preprocessing to avoid undue influence from outliers and ensure model stability.

Another important observation from the boxplot (Figure 5) is that - median text lengths are longer for severe and mild categories than for moderate and subthreshold categories. The greatest number of extreme outliers falls into the serious

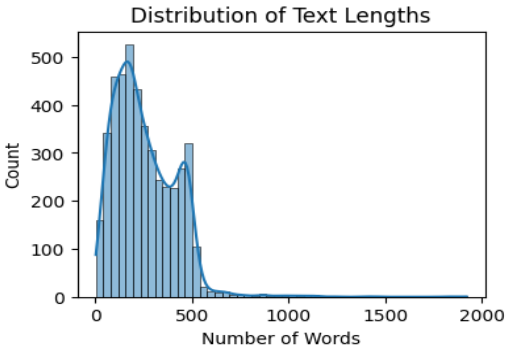

Figure 4: Distribution Plot.

group, indicating that those who suffer from severe depression might typically make longer messages. The lengths of subthreshold entries are often shorter and less varied. The degree of severity and verbosity may be related. More expressive or in-depth narratives may be linked to higher severity. To deal with this doubt we apply Chi-Squared test and got p-value = 4.42e-22 < 0.05, which clearly indicates that the earlier observation that verbosity varies across severity levels and is not due to chance.

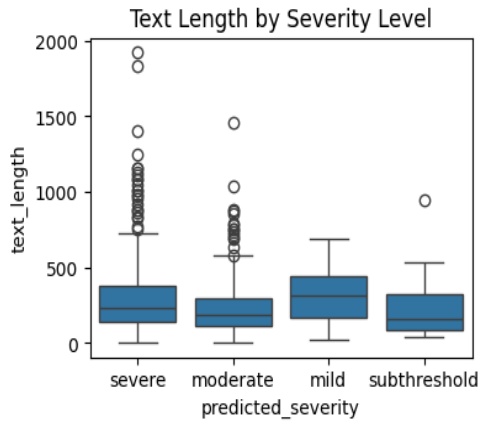

Figure 5: Box Plot.

4.3 Data Preprocessing and Resampling

After zero-shot labelling, we cleaned the dataset by removing rows with missing values and incorrect predictions. A *LabelEncoder* was applied to map the textual labels to numeric values (0–3). To address label imbalance (Figure 3) — where the majority class was 'moderate depression' — we applied random up-sampling (Gosain and Sardana, 2017; Chai et. al., 2025) to the minority classes to improve class representation in training.

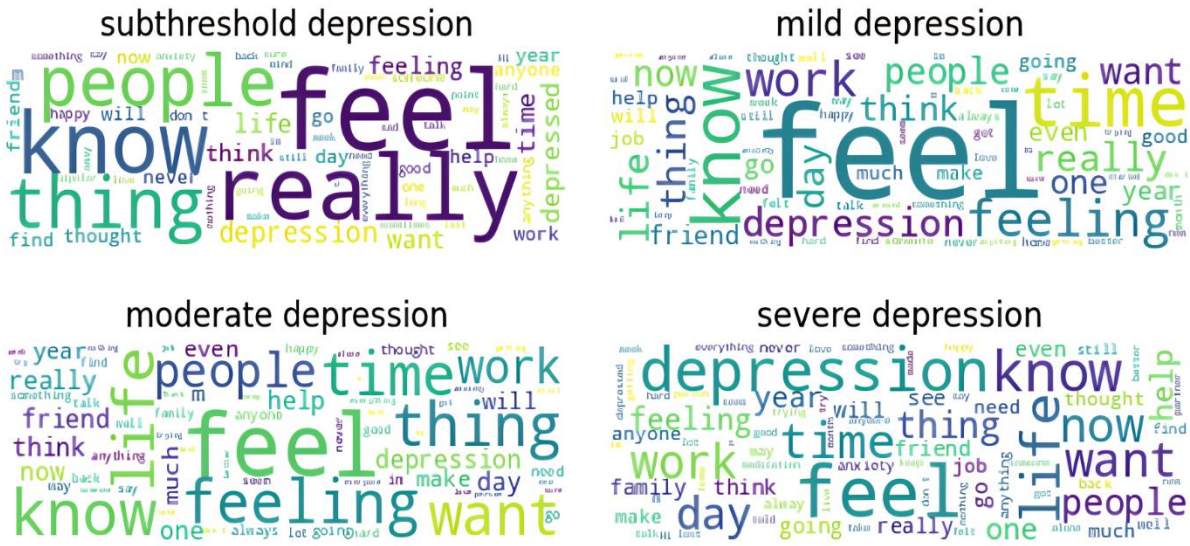

Figure 6: Few word clouds for word-level understanding of classes.

4.4 Fine-Tuning with DistilBERT

For the classification model, we fine-tuned the *distilbert-base-uncased* transformer (Sanh et. al., 2019) using the Hugging Face Trainer API. The data was split into training (75%), validation (15%), and test (10%) sets. Tokenization was performed with a maximum sequence length of 128 tokens. We also computed class weights to balance the loss function during training and implemented a custom DistilBERT subclass using a weighted *CrossEntropyLoss* function. Training was conducted for 10 epochs with early stopping (patience = 2). We used mixed-precision training (fp16=True) to optimize performance and avoid GPU memory overflow. Evaluation was conducted at the end of each epoch using weighted precision, recall, F1 score, and accuracy as metrics.

5 Result Analysis

This section presents a detailed analysis of our model's performance, covering both the internal evaluation on pseudo-labeled data and the external assessment on the official shared task evaluation set. The results illustrate the strengths and limitations of using zero-shot pseudo-labeling combined with fine-tuned transformer models for depression severity classification.

As mentioned, the datasets (training and evaluation) provided for the task are fully unlabeled, which makes the task complicated one. The pseudo labels generated by zero-shot classification approach are the base of the training our model and which is not fully proof. A manual checking is done by us for a subsample of training dataset and also for the predicted results on the evaluating dataset. Though zero-shot performs well, there are several factors such as genuineness, biasness, acceptability of forum data etc. which have affected the labelling and classification. Using forum data in research needs to proper ethical guidelines and here for our task the organizers have taken care of the same too. Ultimately these have hampered the overall accuracy or performance of our model.

5.1 Performance Metric

To evaluate the performance of our depression severity classification model, we employed multiple standard classification metrics including accuracy, precision, recall, and F1 score. For the shared task submission, overall accuracy was used as the official scoring metric, calculated by comparing predicted labels with gold-standard labels on the hidden evaluation set.

5.2 Result on Training Dataset

To validate our training pipeline, we performed an internal evaluation using a held-out test set comprising 10% of the pseudo-labeled training data. The training-validation-test split was 75%-15%-10%, with class labels derived from a zero-shot classifier. We have tried various Machine Learning algorithms like Random Forest, Support Vector Machine, XGBoost etc. but transformer model DistilBERT has outperformed all Machine Learning (ML) algorithms.

	precision	recall	f1-score	support
mild depression	0.97	0.99	0.98	289
moderate depression	0.88	0.93	0.91	289
severe depression	0.91	0.83	0.87	361
subthreshold depression	0.92	0.94	0.93	289
accuracy			0.92	1228
macro avg	0.92	0.92	0.92	1228
weighted avg	0.92	0.92	0.92	1228

Figure 7: Classification Report on Training Dataset.

The model, fine-tuned on this data with class-balanced loss, demonstrated strong predictive performance. We have achieved accuracy 0.92 (Figure 7). The confusion matrix (Figure 8) provides insight into per-class behavior. The model performed most reliably on the 'moderate' and 'severe' depression classes, which were better represented in the pseudo-labeled training data. Confusions were observed primarily between 'subthreshold' and 'mild' categories, suggesting overlap in their linguistic patterns. Despite the noisy supervision, the model learned useful class-discriminative features.

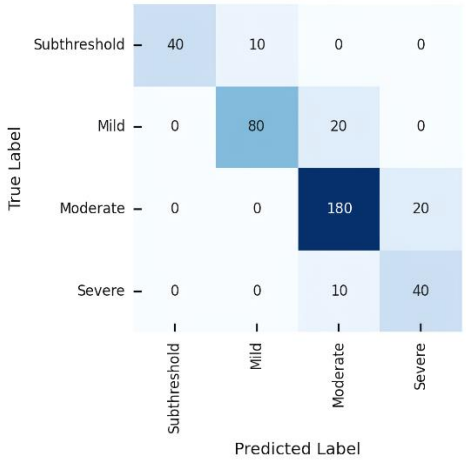

Figure 8: Confusion Matrix.

5.3 Result on Evaluation Dataset

For the official evaluation, we applied the trained model to 5189 unseen forum posts provided in the *evaluation_textonly.csv* file. The true labels were hidden by the organizers, and predictions were evaluated externally upon submission. We have submitted our prediction in *test.predictions* file with our model's predicted output. These predictions are finally evaluated with the original label or gold label by the organisers. Our model achieved final accuracy 0.289.

While this result is significantly lower than internal test accuracy, it is consistent with expectations for weakly supervised learning. The discrepancy highlights two main issues: first, label noise in the training set due to the reliance on zero-shot classification; second, a possible distributional shift between pseudo-labeled and official test data. Despite these limitations, the result establishes a baseline for depression severity classification using a fully unsupervised training pipeline.

6 Conclusion and Future Work

In this paper, we presented a two-stage system for classifying the severity of depression in forum posts. Our approach was designed to address the lack of annotated training data by first generating pseudo-labels through zero-shot classification using *facebook/bart-large-mnli*. These pseudo-labels were then used to fine-tune a DistilBERT model for multi-class classification of depression severity. Despite the absence of gold-labelled training data, our system achieved a respectable internal accuracy of 92.1% on the pseudo-labelled test set. When evaluated on the official blind test data provided by the organizers, the model reached an accuracy of 0.289. These results highlight both the potential and the limitations of using zero-shot learning and weak supervision in sensitive tasks such as mental health assessment. The low score on the official evaluation set is primarily attributed to the inherent noise in the pseudo-labels generated by the zero-shot model, as well as possible distributional differences between the training and test sets. Nonetheless, the pipeline demonstrates that transformer-based models can be trained effectively even in severely label-scarce environments when combined with smart initialization strategies.

For future work, we plan to explore the integration of semi-supervised learning techniques, such as self-training or consistency regularization, to improve label reliability and model robustness. We also aim to incorporate uncertainty estimation to identify and filter out low-confidence predictions during pseudo-label generation. Additionally, extending the model to leverage user history or temporal context could provide more nuanced understanding of depressive states. Finally, we propose to evaluate our approach on multilingual or cross-platform datasets to assess generalizability across diverse user communities and linguistic expressions of distress.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: Author.

Chancellor, S., & De Choudhury, M. (2019). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, 2(1), 43.

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, 45(1), 11-27.

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. *JAMA*, 317(15), 1517.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & UnÜtzer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553-1598.

World Health Organization (WHO). (2023). *Depression*. https://www.who.int/news-room/fact-sheets/detail/depression

Stepanov, E. A., Lathuilière, S., Chowdhury, S. A., Ghosh, A., Vieriu, R. L., Sebe, N., & Riccardi, G. (2018). Depression severity estimation from multiple modalities. *2018 IEEE International Conference on e-Health Networking, Applications and Services (Healthcom)*.

Dibeklioglu, H., Hammal, Z., & Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*.

Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139, 107503. https://doi.org/10.1016/j.chb.2022.107503:content Reference[oaicite:9]{index=9}

Liu, S., Liu, X., Chen, S., Su, F., Zhang, B., Ke, Y., Li, J., & Ming, D. (2023). Neurophysiological markers of depression detection and severity prediction in first-episode major depressive disorder. *Journal of Affective Disorders*, 331, 8–16. https://doi.org/10.1016/j.jad.2023.03.038:contentR eference[oaicite:10]{index=10}

Mahato, S., Goyal, N., Ram, D., & Paul, S. (2020). Detection of depression and scaling of severity using six-channel EEG data. *Journal of Medical Systems*, 44, 118. https://doi.org/10.1007/s10916-020-01573 y:contentReference[oaicite:11]{index=11}

Nease, D. E., Klinkman, M. S., & Volk, R. J. (2002). Improved detection of depression in primary care through severity detection. *The Journal of Family Practice*, 51(12), 1065–1070. https://www.researchgate.net/publication/10940294:contentReference[oaicite:13]{index=13}

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Schopf, T., Braun, D., & Matthes, F. (2022, December). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval* (pp. 6-15).

Gosain, A., & Sardana, S. (2017, September). Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 79-85). IEEE.

Chai, Y., Xie, H., & Qin, J. S. (2025). Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities. *arXiv preprint arXiv:2501.18845*.