Comparative Analysis of Neural Network Architectures for Classifying Depressive Content in Social Networks

Yntymak Abdrazakh[1], Rita Ismailova[2], Nurseit Zhunissov[3]*,
Arypzhan Aben[4], Anuarbek Amanov[5], Aigerim Baimakhanova[6]*
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[1, 3, 4, 5, 6]
Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan[2]

***Abstract*—Depression-related language on social media provides measurable signals for population-level mental-health research, yet model selection remains sensitive to evaluation protocol, domain shift, class imbalance, and computational constraints. This study benchmarks CNN, LSTM, and transformer encoders (BERT, RoBERTa, DistilBERT, and MentalBERT) for binary depression-indicative versus control classification on a unified corpus of 19,800 English posts/comments aggregated from three platforms (Reddit, Twitter, and Facebook) under a consistent preprocessing pipeline. We report two complementary evaluation protocols: (1) a fixed-split single-run baseline for a comparable snapshot, and (2) a five-seed repeated-run protocol with statistical testing (effect sizes and multiple-comparison correction) to quantify variability and reduce sensitivity to initialization effects. Under repeated-run reporting, MentalBERT achieves the best overall performance (F1 = 0.918 ± 0.005; AUC = 0.962 ± 0.002), while CNN/LSTM baselines show lower robustness under cross-platform transfer. Cross-domain experiments reveal a consistent performance drop relative to in-domain evaluation, confirming non-trivial platform shift and motivating robustness-aware reporting for deployment-oriented settings. In addition to predictive metrics, we report training time, inference latency, and derived throughput to support practical model selection for use cases such as moderation pipelines and screening/triage dashboards.**

***Keywords*—Depression detection; social media text; natural language processing; cross-platform evaluation; robustness; statistical significance testing; transformer-based models; CNN; LSTM; BERT; MentalBERT**

I. Introduction

In modern society, depression has become an increasingly pressing public-health challenge, influencing not only individual well-being but also broader social and economic outcomes. According to the World Health Organization, depression affects approximately 332 million people worldwide (about 4% of the global population) and is a common mental disorder that can substantially impair daily functioning and, in severe cases, increase suicide risk [1]. In parallel, the rapid growth of digital footprints on social media platforms (e.g., Twitter, Reddit, Facebook) provides large-scale textual evidence where users may express emotional distress and related indicators, making text-based analysis a promising direction for population-level risk monitoring and research support. Prior work shows that linguistic signals in social-media text are associated with depression-related outcomes and can support NLP-based screening and monitoring in research settings [2], [3].

In recent years, there has been substantial progress in automated assessment of a person's psychological state through text analysis. Natural Language Processing (NLP), combined with machine-learning methods, provides computational tools for detecting depression-related signals from user-generated statements and other online textual traces [4]. Recent surveys and systematic reviews describe the evolution of this field from feature-engineered classifiers toward deep representation learning, including neural architectures and transformer-based approaches for text classification [5], [6]. In this context, NLP-based methods can support research-oriented screening and monitoring of depression and anxiety indicators; however, model outputs should be interpreted as risk signals rather than clinical diagnoses.

Early computational mental-health studies on online platforms often relied on bag-of-words or lexicon-based representations combined with linear classifiers (e.g., SVM, Naïve Bayes, logistic regression) as strong baselines for depression-related text classification. In community and forum settings, such approaches were evaluated on large-scale resources built from users' self-reported signals (e.g., self-reported diagnosis–based forum resources), which helped establish benchmark tasks for depression detection from language [7]. However, feature-engineered pipelines can struggle with subtle semantics, pragmatic cues, and longer-range context in noisy user-generated text, motivating deep representation learning. This shift promoted architectures that learn hierarchical and sequential representations directly from text, including CNN- and LSTM-based models that became strong baselines for sentence- and post-level classification tasks [8],[9]. More recently, transformer-based encoders and ensemble strategies have further improved contextual modeling for depression detection across social platforms, reinforcing the move toward attention-based architectures in modern benchmarks [10].

Transformer-based architectures have substantially advanced NLP by enabling global context modeling via self-attention mechanisms [11]. Building on this framework, BERT introduced deep bidirectional contextual representations and achieved strong performance across a wide range of language understanding and classification tasks, including text

*Corresponding author.

classification settings relevant to affective and mental-health NLP [12].

The present study provides a comparative assessment of CNN, LSTM, and BERT models in the context of depression detection from text-based data. All models were trained under identical conditions to ensure comparability. The analysis focuses on key performance metrics such as Accuracy, Precision, Recall, F1-score, and AUC, aiming to determine which architecture demonstrates the highest reliability in identifying depressive language patterns.

Despite the strong performance of CNN and LSTM baselines, recent work increasingly favors transformer-based encoders, particularly BERT and its variants, for mental-health–related text classification, due to their ability to model long-range context and bidirectional dependencies [12], [13]. More recently, several works have explored fine-tuning large language models for depression detection on social-media benchmarks and reported strong performance under their experimental settings [14]. In broader benchmarking and comparative studies, transformer-based encoders are frequently reported to be competitive on mental-health NLP tasks, including depression and anxiety detection [15]. However, statistical significance testing is sometimes omitted or misapplied in NLP benchmarking, which can lead to overconfident conclusions [16]. Practical deployment must still consider computational requirements, inference latency, and memory footprint compared to lighter neural architectures [17].

Despite substantial progress in neural text classification for mental health applications, the question of how different deep learning architectures compare under controlled and deployment-relevant conditions remains insufficiently explored. Existing studies often focus on a single model family, report results on a single dataset, or rely on one-off train–test splits without accounting for variability induced by random initialization and training dynamics. As noted in prior surveys and comparative reviews, comprehensive analyses that systematically contrast CNN, LSTM, and transformer-based models in depression-related text classification remain relatively scarce [15]. As a result, model selection is frequently based on isolated performance snapshots rather than on robust, reproducible evidence, complicating the identification of architectures that achieve a balanced trade-off between predictive accuracy and computational efficiency in real-world systems.

The purpose of this study is to conduct an empirical comparison of CNN, LSTM, and BERT architectures for binary classification of social media texts with respect to depressive versus non-depressive language. In addition to standard quality metrics (Accuracy, Precision, Recall, F1-score, and AUC), we quantify run-to-run variability under repeated training, report effect sizes with multiple-comparison control, and evaluate cross-platform transfer to better reflect deployment-oriented conditions.

Deployment motivation. The intended deployment setting is a post-level text classifier integrated into a decision-support workflow (e.g., a research screening dashboard or a content-moderation triage pipeline) where the model flags potentially depressive posts for human review. In such settings, the objective is not clinical diagnosis but reliable risk-sensitive prioritization under platform shift, bounded compute, and real-time throughput constraints.

In summary, this work (1) performs a controlled comparison of CNN, LSTM, and multiple transformer encoders under identical preprocessing and evaluation conditions; (2) evaluates both a single-run baseline and a repeated-run protocol with statistical testing to account for performance variability; and (3) complements predictive metrics with robustness (cross-platform transfer) and efficiency indicators (training time and inference latency) to support deployment-oriented model selection.

This study hypothesises that transformer-based encoders will outperform lightweight CNN/LSTM baselines on depression-indicative text classification because they model non-local context and subtle semantics. Although this direction is broadly expected, its magnitude, stability across random seeds, and robustness under cross-platform shift are not guaranteed and are often underreported in single-split comparisons. Therefore, the main novelty of this work is not architectural invention but a deployment-relevant evaluation design: we jointly analyze baseline accuracy, repeated-run stability, cross-platform degradation, and efficiency indicators under a unified preprocessing and training protocol.

The remainder of the study is organized as follows. Section II reviews related work and highlights methodological gaps in prior comparative studies. Section III describes the datasets, preprocessing pipeline, evaluation protocols, and statistical testing. Section IV reports experimental results, including baseline comparisons, transformer variants, and cross-platform transfer. Section V discusses deployment implications, limitations, and future research directions. Section VI concludes the study.

II. Related Works

Research on depression detection from text has evolved from feature-engineered machine-learning pipelines toward deep representation learning. Early studies relied on linear classifiers and handcrafted lexical and psycholinguistic features, but such approaches often struggled to capture compositional meaning and longer-range context in mental-health–related language, especially in noisy, user-generated social-media posts where context and pragmatics are hard to capture [4]. As deep learning became standard, neural baselines such as CNN and LSTM gained popularity due to their ability to learn task-relevant representations directly from data. CNN-based models are effective at extracting local n-gram patterns for sentence- and post-level classification [8], while recurrent architectures such as LSTM better capture sequential dependencies that may reflect evolving emotional framing in text [9].

Critical synthesis and gap. Many prior depression-detection comparisons are difficult to reconcile because they (1) use different label definitions (self-report vs symptom proxy), (2) mix platform-specific preprocessing and splits that may leak user timelines, (3) report a single train–test split without seed variability, and (4) omit efficiency reporting beyond point accuracy. Consequently, it is often unclear whether reported gains reflect model capability, data preprocessing choices, or favorable randomness. Domain-adaptive transformers such as

MentalBERT are frequently reported to outperform general-purpose BERT in mental-health language, but the magnitude varies across datasets. In our controlled benchmark, MentalBERT improves mean F1 by approximately 0.052 over standard BERT under Protocol B (Table IV), supporting the practical value of domain-adaptive pretraining when computational budgets permit. In addition, statistical testing protocols are often underspecified in NLP comparisons [16].

Comparative studies and systematic reviews often report that transformer-based encoders are competitive or superior to classical baselines on mental-health text classification benchmarks, although results depend on dataset construction and evaluation protocol [15]. At the same time, practical deployment must account for computational and memory costs of attention-based models, which motivates efficiency-oriented transformer variants and compact encoders when latency or footprint is constrained [17]. In addition, domain-adaptive pretraining (e.g., MentalBERT) aims to improve sensitivity to mental-health–specific language patterns and implicit distress markers. More recently, several works also explored fine-tuning large language models for depression detection, reporting strong performance on selected social-media benchmarks [14].

A complementary research direction focuses on domain-adaptive pretraining. MentalBERT is a representative example of a transformer pretrained on mental-health–related corpora (including content from mental-health communities), to increase sensitivity to psychologically salient cues and implicit distress markers. Such domain adaptation has been reported to improve performance over general-purpose encoders in settings where the target language includes community-specific vocabulary, self-referential statements, and non-standard phrasing. Beyond domain adaptation, integrative reviews emphasize that affective expression and linguistic markers in social-media text are consistently associated with mental-health–related outcomes, supporting emotion-aware feature analysis in computational mental-health research [18].

Datasets and evaluation campaigns have also expanded substantially. Classic corpora such as RSDD and platform-specific datasets enabled large-scale depression-related text classification and benchmarking [7], [19]. More recent resources emphasize richer supervision (e.g., DSM-oriented labeling), multi-task objectives, and evaluation designs that better reflect temporal dynamics and domain shift. For example, MentalHelp provides a multi-task dataset for mental-health objectives in social media [20], while ReDSM5 introduces depression detection labels aligned with DSM-5-oriented criteria on Reddit content [21]. In addition, user-level and multimodal datasets have been proposed to capture changes in behavior and language over time, enabling evaluation beyond isolated posts [22]. Community-driven shared tasks such as CLPsych encourage standardized protocols that emphasize robustness and longitudinal modeling in realistic scenarios [23]. Expanding beyond English and beyond a single platform is also an active direction; for instance, Arabic mental-health resources such as CARMA support cross-lingual and cross-cultural studies [24].

Overall, prior work indicates that realistic deployment-oriented evaluation should consider not only in-domain accuracy but also robustness under repeated runs, cross-platform transfer, and computational constraints. However, controlled comparisons that jointly analyze predictive quality, repeated-run stability, cross-platform robustness, and practical efficiency indicators under a unified pipeline remain limited [17]. Moreover, many studies report single-split results without repeated-run variability, which complicates robust model selection for deployment-oriented settings. This motivates the comparative and deployment-oriented analysis performed in the present study, as well as the consideration of early-risk prediction setups such as CLEF eRisk, which emphasize longitudinal evidence and early detection under constrained information [25].

III. Materials and Methods

A. Data

To conduct a comparative analysis, we relied on publicly available research corpora originating from multiple social media platforms, which helps capture variation in writing style, message length, and the way depressive states may be expressed in different online communities. The following datasets were used:

1) Reddit Self-reported Depression Diagnosis (RSDD): A user-level corpus derived from Reddit timelines where users self-report a depression diagnosis, paired with matched control users. The original resource covers posts from approximately 9,000 diagnosed users (with a larger set of matched controls) [7]. For the present experiments, we derive a post-level binary classification subset by sampling posts from the corresponding user timelines under a fixed protocol and applying filtering and normalization. Because the source dataset is user-level, we enforce user-disjoint train/validation/test splits to prevent leakage, ensuring that posts from the same user appear in only one split.

2) Twitter-based depression dataset (Twitter-STMHD): A user-level benchmark resource for mental-health analysis on Twitter [19]. From this dataset, we derive a post-level experimental subset of approximately 6,500 English tweets for binary depression-indicative versus control classification after filtering and normalization. As Twitter-STMHD is provided at the user level, we construct the post-level subset by sampling tweets from user timelines while enforcing user-disjoint splits to prevent user overlap across train, validation, and test.

3) Facebook comments corpus: Following Islam et al. [26], we use a Facebook comments corpus derived from publicly accessible pages and labeled into two categories (depression-indicative vs non-depressive) under the authors' labeling procedure. The source study reports 7,145 labeled comments (4,149 depression-indicative; 2,996 non-depressive). After applying our preprocessing and filtering pipeline (e.g., removing empty/duplicate entries and non-English fragments), we retain 4,300 instances for controlled experiments while preserving the original label mapping. In addition, to keep the merged corpus prevalence consistent with the overall non-depressive majority observed across sources (Table I), we subsampled the retained Facebook data without changing label definitions. Due to platform policy and data-sharing

constraints, we report only aggregated results and do not redistribute raw Facebook text.

Overall, the unified dataset comprises 19,800 post-level instances spanning all three platforms (9,000 Reddit posts, 6,500 tweets, and 4,300 Facebook comments). We use a fixed stratified split into training, validation, and test sets (80%/10%/10%). For user-level sources (RSDD and Twitter-STMHD), splits are user-disjoint to avoid leakage across a user's timeline; for Facebook, comments are split at the instance level while preserving class balance.

TABLE I. PLATFORM-WISE CLASS DISTRIBUTION AFTER PREPROCESSING

Platform	Total (N)	Depressive, n (%)	Non-depressive, n (%)
Reddit	9,000	3,600 (40.0%)	5,400 (60.0%)
Twitter	6,500	2,400 (36.9%)	4,100 (63.1%)
Facebook	4,300	1,800 (41.9%)	2,500 (58.1%)
Total (all platforms)	**19,800**	**7,800 (39.4%)**	**12,000 (60.6%)**

Across platforms, depressive posts are less frequent overall (Table I). We tested whether class proportions differ across platforms using a chi-square test of independence and found statistically significant, but small, differences ($\chi^2(2) = 28.96$, $p = 5.14e\text{-}07$; Cramér's $V = 0.038$). These supports treating platform shift as a meaningful factor when interpreting cross-platform transfer results.

The merged corpus exhibited a moderate skew toward non-depressive posts. To mitigate potential bias during training without distorting evaluation, we applied a conservative balancing strategy after performing the train/validation/test split and only on the training split. Specifically, we used mild random under sampling of the majority class and semantic-preserving text augmentation for the minority (depressive) class, while keeping the validation and test splits unchanged. This approach increases exposure to depressive patterns during learning while preserving a realistic evaluation setting and preventing leakage from synthetic text into evaluation.

All datasets used in this study are publicly available research corpora or publicly accessible social-media text collections described by prior work and used under their respective terms and platform constraints. In our local processing, we removed or masked user handles, mentions, URLs, and other explicit identifiers where present, and we used the data strictly for aggregated modeling and evaluation rather than any individual-level inference.

B. Text Preprocessing

Social-media text is typically noisy and informal, containing emojis, URLs, mentions, hashtags, non-standard spelling, and grammatical deviations. To reduce noise while preserving emotionally informative patterns, we applied a multi-stage preprocessing pipeline designed for user-generated posts. The process included six interrelated stages (Fig. 1) that transform raw posts (raw post text) into model-ready inputs while preserving linguistically and affectively informative cues.

1) Cleaning: We removed purely technical tokens (URLs, @mentions, and obvious HTML artifacts) and normalized whitespace. Importantly, we preserved punctuation patterns that may convey affect (e.g., ellipses "...", repeated exclamation marks "!!!", and expressive repetition) when they functioned as emotional markers rather than boilerplate. Emoji handling was conservative to avoid discarding affective cues.

2) Tokenization: To ensure comparability across model families, we used word-level tokenization for CNN/LSTM baselines and subword tokenization for transformer encoders using the official pretrained tokenizers (e.g., WordPiece for BERT). This design prevents tokenization mismatch from confounding architecture comparisons.

3) Lemmatization: Lemmatization was applied conservatively with part-of-speech constraints to avoid semantic distortion in colloquial expressions. We avoided aggressive stemming that may alter psychologically relevant wording.

4) Normalization: Text was lowercased to reduce sparsity; however, we avoided aggressive stop-word removal because function words, negations, and stylistic markers can be informative in depression-related language, as emphasized in methodological reviews of mental-health prediction from social media [4]. Normalization was kept light to preserve expressive spelling patterns that may indicate emotional intensity.

5) Padding/truncation: All inputs were padded/truncated to a fixed length of 128 tokens. When truncation was required, we applied a consistent policy across models by retaining the most informative part of the sequence under a fixed-length constraint.

6) Model input: For transformer-based models, we constructed input IDs and attention masks following the standard input formatting of the corresponding pretrained tokenizer, with truncation/padding to a fixed maximum length. For CNN and LSTM, the same cleaned texts were mapped into index sequences using the baseline vocabulary and embeddings pipeline. We fixed the maximum sequence length and padding strategy consistently across transformer models to ensure comparable computational budgets.

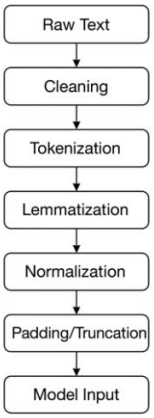

Fig. 1. Algorithmic pipeline for preprocessing textual data before neural classification.

C. Model Architectures

The dataset exhibited moderate class imbalance between depressive and non-depressive posts, which can negatively affect model generalization. Initially, the Synthetic Minority Over-sampling Technique (SMOTE) was considered for balancing the dataset [27]. SMOTE operates by interpolating minority samples in a continuous feature space; however, for sequence-based neural models that require token sequences as inputs, such interpolation does not directly yield new valid text instances. Therefore, we adopted training-only resampling and semantic-preserving text augmentation rather than relying on synthetic oversampling for the main experiments [27], [28], [29].

To mitigate class imbalance without distorting evaluation, we applied a conservative balancing strategy only on the training portion of the data after performing the train/validation/test partitioning. Specifically, we used mild random undersampling of the majority class and semantic-preserving text augmentation for the minority (depressive) class, while keeping the validation and test splits unchanged. This prevents leakage, preserves realistic prevalence during evaluation, and reduces the risk of over-interpreting a single favorable run. A systematic ablation study comparing multiple augmentation families is beyond the scope of this work and is left for future research.

To classify depressive content, we implemented and compared three widely adopted neural architectures: a convolutional model for capturing local n-gram patterns (CNN) [8], a recurrent model designed to model sequential dependencies (LSTM) [9], and a transformer-based architecture relying on bidirectional contextual encoding (BERT) [12]. Fig. 2 summarizes the compared model families at a high level.

1) Convolutional neural network: A model that captures local n-gram patterns using convolutional filters and pooling. Following common sentence-level CNN setups, we used convolution kernels of sizes 3, 4, and 5 with GlobalMaxPooling, and applied dropout (0.5) for regularization to mitigate overfitting in short-text settings. We additionally align this configuration with depression-oriented CNN pipelines reported in prior studies [30].

2) Long short-term memory: is a recurrent architecture capable of capturing temporal dependencies and context in sequences of words. The model included one bidirectional LSTM layer with a latent state dimension of 128 and an output fully connected layer.

3) Bidirectional encoder representations from transformers: A transformer-based architecture that produces bidirectional contextual representations of tokens using self-attention. In our experiments, we used the pre-trained BERT-base-uncased model and fine-tuned it on the prepared corpus for depression classification, following the standard fine-tuning protocol for BERT-style encoders.

To ensure a fair comparison, we standardized the training pipeline across models and tuned hyperparameters within controlled ranges. For transformer-based models (BERT, RoBERTa, DistilBERT, MentalBERT), we fine-tuned pretrained encoders using AdamW with a linear learning-rate scheduler and searched learning rates in {1e−5, 2e−5, 3e−5, 5e−5} and batch sizes in {16, 32}, with 2–5 epochs and early stopping based on validation loss (patience 2–3). For CNN and LSTM, we tuned the hidden size and dropout rate and selected learning rates via Bayesian optimization to balance convergence stability and generalization. Unless explicitly stated, we kept the maximum sequence length fixed at 128 tokens and used the same data split and preprocessing pipeline across all baseline experiments. This setting follows common practice in transformer fine-tuning for sentence-level classification and provides a strong efficiency–accuracy trade-off, since self-attention has quadratic complexity in sequence length. In particular, the original BERT training strategy used a sequence length of 128 for the majority of training steps to reduce computational cost, increasing to 512 only for a small portion of training.

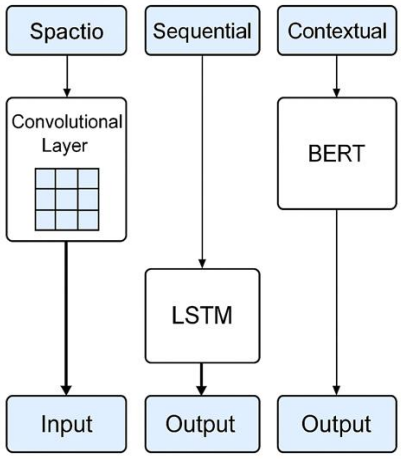

Fig. 2. Comparison of neural network architectures for text classification.

D. Evaluation Metrics

We evaluate all models using standard binary classification metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC. Together, these metrics characterize overall correctness, false-alarm behavior, and the ability to identify depressive content without missing critical cases. In mental health screening settings, Recall is particularly important because false negatives may lead to missed at-risk users, while Precision helps limit unnecessary false alarms. AUC-ROC is additionally reported as a threshold-independent indicator of class separability [31].

- Accuracy (Classification accuracy).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Accuracy represents the proportion of correctly classified samples among all observations. It provides a general measure of model performance but may be misleading for imbalanced datasets.

- Precision (Positive Predictive Value)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Precision indicates the proportion of true positive predictions among all predicted positives. High precision is essential when minimizing false alarms is critical.

- Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

Recall measures the model's ability to identify all relevant positive instances. In mental health detection tasks, high recall is vital to avoid missing cases that require attention.

- F1-score (Harmonic mean of precision and recall)

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

F1-score balances precision and recall, providing a single measure of model effectiveness. It is particularly useful when both false positives and false negatives are costly.

- AUC (Area Under the ROC Curve)

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (5)$$

where:

$TPR = \frac{TP}{TP+FN}$ – True Positive Rate (sensitivity),

$FPR = \frac{FP}{FP+TN}$ – False Positive Rate.

AUC measures the area under the Receiver Operating Characteristic (ROC) curve, showing the model's ability to discriminate between classes. A higher AUC indicates stronger separability.

E. Evaluation Protocols

We report results under two complementary protocols. Protocol A (single-run baseline) reports performance on a fixed held-out split using one training run per model configuration (see Table III and Fig. 3–5). Protocol B (repeated runs) evaluates robustness by training each configuration five times with different random seeds while keeping the same split and preprocessing pipeline fixed; results are summarized as mean ± standard deviation in Table IV. Repeated-seed reporting is important because fine-tuning outcomes can vary due to initialization, minibatch order, and early-stopping dynamics. Therefore, we report seed-averaged performance to reduce the risk of over-interpreting a single favorable run and to provide a more stable basis for model comparison [32], [33]. This separation prevents mixing single-run snapshots with stability-oriented reporting and improves reproducibility.

F. Statistical Analysis

To assess whether observed performance differences between models were statistically meaningful, we tested model comparisons on Protocol B per-run F1 scores (n = 5 seeds per model). We first applied one-way ANOVA over per-run F1 to test the global null hypothesis of equal means. For planned pairwise localization, we then performed paired t-tests on seed-matched runs comparing BERT against each lightweight baseline (CNN and LSTM). To control family-wise error under these two planned comparisons, we applied Holm's step-down adjustment. Alongside p-values, we report paired effect sizes using Cohen's d_z, computed as the mean of paired differences divided by the standard deviation of paired differences. As a non-parametric sensitivity check for small n, we additionally report Wilcoxon signed-rank tests [34], but we interpret them cautiously because n = 5 offers limited power.

IV. Experimental Results

A. Data Preparation

For the experiments, we used the combined corpus described in the "Data" subsection of MATERIALS AND METHODS. It consists of curated subsets from Reddit, Twitter, and Facebook datasets and totals approximately 19,800 texts after preprocessing and deduplication. The detailed dataset sources, subset sizes, and selection criteria are provided in the main "Data" description to ensure clarity and reproducibility.

B. Model Training

Table II summarizes the training setup and computational characteristics of the baseline architectures under identical preprocessing and a fixed train/test split. CNN and LSTM were optimized using Adam, while transformer-based encoders were fine-tuned using AdamW with a linear learning-rate scheduler, following standard practice for BERT-style models. All models used binary cross-entropy loss for the binary classification objective. All experiments were conducted in Google Colab using an NVIDIA Tesla T4 GPU. We recorded the runtime environment (Python, CUDA, PyTorch/TensorFlow, NumPy/SciPy, scikit-learn, and Hugging Face Transformers) and the full configuration (random seeds, batch size, learning rate, early stopping, and hyperparameter ranges) to support reproducibility.

TABLE II. Training Setup and Computational Characteristics of the Baseline Architectures

Model	Framework	Key Parameters	Training Time (GPU)
CNN	TensorFlow/Keras	128 filters, kernel={3,4,5}, dropout=0.3	~0.3 h
LSTM	TensorFlow/Keras	128 units, dropout=0.3	~0.4 h
BERT (base-uncased)	Hugging Face Transformers	lr=2e-5, batch=16, epochs=3	~3.4 h

C. Results and Comparison

After training, each model's performance was evaluated using standard classification metrics (Accuracy, Precision, Recall, F1-score, and AUC). The results of calculating metrics for each architecture are shown in Table III.

TABLE III. Comparison of Classification Metrics Across Baseline Models

Metric	CNN	LSTM	BERT
Accuracy	0.820	0.870	**0.897**
Precision	0.800	0.860	**0.887**
Recall	0.790	0.880	**0.892**
F1-score	0.795	0.870	**0.889**
AUC	0.850	0.910	**0.943**

Analysis of the table shows that all three architectures have demonstrated stable results, but performance differs across architectures in a consistent manner.

The CNN model showed satisfactory Accuracy (0.820) and AUC (0.850) values, which confirms its ability to isolate local text patterns and n-grams. However, the lower Recall (0.790) and F1-score (0.795) scores indicate that CNN has limited capabilities in recognizing contextual dependencies characteristic of natural language.

The LSTM model, due to its recurrent structure, demonstrated a deeper understanding of sequential dependencies and improved all metrics compared to CNN. In particular, Recall increased to 0.880, and AUC increased to 0.910, which indicates a higher ability of the model to detect depressive texts without a significant increase in false positives.

The BERT model showed the highest results, reaching the maximum values Accuracy = 0.897, Recall = 0.892, F1 = 0.889 and AUC = 0.943. This result is explained by the use of the self-attention mechanism, which allows the model to take into account the context of the entire sentence, not just neighboring words. This is especially important when analyzing emotionally laden statements, where the meaning may depend on complex contextual relationships.

A high Recall value (0.892) indicates the minimum number of missed cases of depressive texts, which is of key importance for monitoring the psycho-emotional state of users on social networks. Concurrently, high Precision (0.887) indicates a low level of false alarms, which makes the BERT model the most balanced and reliable solution among the tested architectures.

BERT surpasses the traditional CNN and LSTM architectures in all major metrics, especially in F1 and AUC, which confirms its high ability for contextual text analysis.

LSTM remains an acceptable option for scenarios with limited computing resources, providing a good balance of accuracy and speed.

CNN, despite its relatively low completeness (Recall), can be used for basic or mobile applications where speed and ease of implementation are critical.

Thus, the BERT model has demonstrated the best balance between accuracy, completeness, and the ability to distinguish classes, which makes it preferable for practical systems for analyzing depressive content on social networks.

Fig. 3–5 provide complementary visual evidence for the baseline comparison in Table III, including a metric profile (Fig. 3), ROC curves across thresholds (Fig. 4), and confusion matrices that summarize error types (Fig. 5).

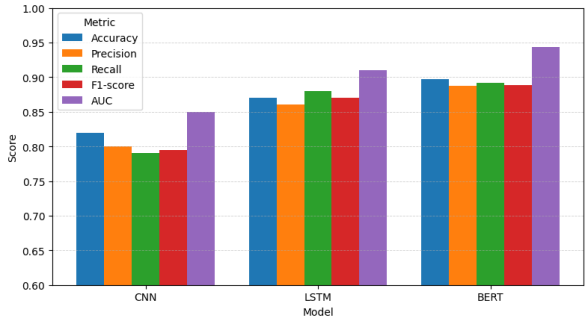

Fig. 3. Comparison of classification metrics across models.

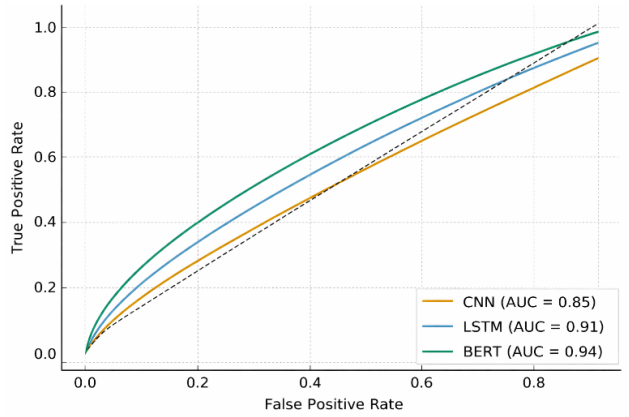

Fig. 4. ROC curves for each architecture.

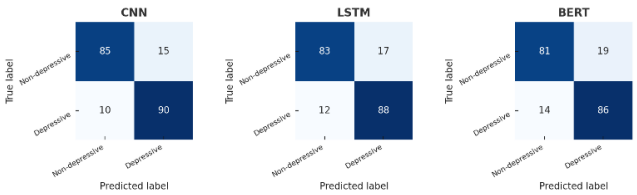

Fig. 5. Confusion matrices illustrating classification performance.

D. Statistical Analysis

To assess whether the observed differences in predictive quality are statistically meaningful, we performed significance testing on per-run F1 scores obtained under Protocol B (five independent runs with different random seeds, n = 5 per model). We first applied one-way ANOVA across the three baseline models (CNN, LSTM, BERT) and obtained $F(2, 12) = 33.73$, $p = 1.19e-05$, indicating that at least one model differs in mean F1 under repeated training.

To localize pairwise differences while limiting multiple comparisons, we performed planned post-hoc paired t-tests comparing BERT to each lightweight baseline using seed-matched runs, reporting Holm-adjusted p-values and paired effect sizes (Cohen's d_z). BERT outperforms CNN ($p = 0.004$; Holm-adjusted $p = 0.008$; $d_z = 2.64$) and outperforms LSTM ($p = 0.017$; Holm-adjusted $p = 0.017$; $d_z = 1.76$). In addition, LSTM outperforms CNN in these runs ($p = 0.003$; $d_z = 3.00$). Overall, the results suggest that the performance advantages of contextual modeling are unlikely to be explained solely by random initialization under the evaluated setting.

E. Summary of Findings

In the baseline comparison, transformer-based encoding (BERT) achieved the most favorable overall metric profile, including a strong F1-score and AUC, indicating improved class separability in depression-related text.

LSTM provided a competitive non-transformer baseline and generally outperformed CNN, consistent with improved modeling of sequential context and longer dependencies.

CNN remained computationally efficient but showed weaker performance on metrics that are sensitive to subtle semantics, suggesting limitations for nuanced mental-health language.

Statistical testing on repeated runs (Protocol B) indicates that the performance differences involving BERT are unlikely to be due to random variation alone, supporting the robustness of the observed ranking.

F. Enabling and Testing New Transformers: RoBERTa, DistilBERT, MentalBERT

To extend the baseline comparison beyond CNN/LSTM/BERT, we evaluated additional transformer-family encoders that represent different trade-offs in pretraining objectives, parameter efficiency, and domain adaptation. Specifically, we considered RoBERTa, which improves upon BERT-style pretraining by removing the next-sentence prediction objective and leveraging larger training corpora [36]; DistilBERT, a compressed variant designed to reduce computation with minimal loss in accuracy [37]; and MentalBERT, a domain-adapted transformer pretrained on mental-health–related corpora, aimed at improving sensitivity to psychologically salient cues [13]. Importantly, MentalBERT retains the BERT-base architecture and WordPiece tokenization; thus, improvements are attributable to domain-adaptive pretraining rather than tokenization changes. A detailed embedding-space similarity analysis (e.g., representation drift between BERT and MentalBERT) is left for future work.

Results are summarized in Table IV under Protocol B, where each model configuration is trained five times with different random seeds on the same fixed split, and the reported values correspond to mean ± standard deviation across runs. For clarity, Table III presents Protocol A as a single-run baseline on the same split; therefore, Table III provides a point-performance snapshot rather than variability estimates. Accordingly, mean ± standard deviation is reported only for Protocol B (Table IV), while Protocol A (Table III) is presented as a single-run baseline snapshot. This two-protocol design complements the single-run snapshot (Protocol A) with variability-aware evaluation (Protocol B), helping characterize stability and reduce sensitivity to initialization effects.

All classification metrics in Tables III and IV are reported in fractional form (0-1). Percent values are used only where explicitly indicated (e.g., Table V). Table IV presents a comparative view of predictive quality (Accuracy, Precision, Recall, F1-score, AUC) together with practical deployment indicators (training time, inference latency, and model size). Overall, RoBERTa demonstrates competitive performance among general-purpose transformer variants, consistent with stronger pretraining on larger corpora. DistilBERT typically shows a small reduction in predictive quality while offering noticeable gains in computational efficiency, making it a reasonable candidate for resource-constrained scenarios. MentalBERT achieves strong performance in this setting, which aligns with the expected benefit of domain-adaptive pretraining for mental-health language, where subtle cues, slang, and indirect phrasing can be critical for correct classification.

Batching throughput note. Let ℓ denote the batch=1 inference latency (ms/sample) reported in Table IV. For a batch size b, the achievable throughput lies between a conservative lower bound of $1000/\ell$ samples/s (no batching speedup) and an optimistic upper bound of $b \cdot 1000/\ell$ samples/s (ideal batching), with actual throughput depending on GPU memory bandwidth, kernel fusion, and framework overheads. Using the Table IV latencies, the corresponding ranges are: BERT (baseline): b=1 59.5; b=8 59.5–476.2; b=16 59.5–952.4; b=32 59.5–1904.8 samples/s. RoBERTa-base: b=1 54.9; b=8 54.9–439.6; b=16 54.9–879.1; b=32 54.9–1758.2 samples/s. DistilBERT: b=1 106.4; b=8 106.4–851.1; b=16 106.4–1702.1; b=32 106.4–3404.3 samples/s. MentalBERT: b=1 50.8; b=8 50.8–406.1; b=16 50.8–812.2; b=32 50.8–1624.4 samples/s.

Efficiency note (throughput and FLOPs). Using the reported batch-1 inference latency, the corresponding batch-1 throughput is approximately 59.5 samples/s (BERT), 54.9 samples/s (RoBERTa), 106.4 samples/s (DistilBERT), and 50.8 samples/s (MentalBERT). Throughput typically increases under batching up to hardware saturation, but the exact gain is device- and implementation-dependent; therefore, we report latency and derived batch-1 throughput as comparable baselines. We also report approximate per-sample forward-pass compute for transformers at sequence length 128 using standard FLOP accounting: a 12-layer BERT/RoBERTa/MentalBERT encoder is on the order of ~22 GFLOPs per sample, while DistilBERT (6 layers) is ~11 GFLOPs. These estimates help explain why DistilBERT provides a favorable efficiency–accuracy trade-off, whereas domain-adaptive models may be preferred when accuracy dominates and compute budgets allow. At small batch sizes, runtime can also be influenced by memory bandwidth and kernel launch overheads; hence, FLOP estimates should be interpreted as approximate compute proxies rather than exact predictors of latency.

G. Qualitative Error Analysis (False Positive / False Negative)

We performed a qualitative inspection of misclassified examples to identify common sources of false positives (FP) and false negatives (FN) across architectures. A recurring pattern for FP errors is the presence of negatively valenced language that expresses temporary frustration or situational stress without clear indicators of persistent depressive symptoms. In such cases, strong sentiment words dominate the surface signal, while the broader self-referential and persistent framing that is more typical of depressive discourse is absent.

TABLE IV. PERFORMANCE COMPARISON OF TRANSFORMER-BASED MODELS FOR DEPRESSIVE CONTENT CLASSIFICATION

Model	Accuracy	Precision	Recall	F1-score	AUC	Training time (hrs)	Inference latency (ms/sample)	Model size (MB)
BERT (baseline)	0.897 ± 0.004	0.887 ± 0.004	0.892 ± 0.004	0.889 ± 0.004	0.943 ± 0.003	3.4 ± 0.2	16.8 ± 0.5	418
RoBERTa-base	0.912 ± 0.003	0.904 ± 0.003	0.917 ± 0.003	0.910 ± 0.003	0.956 ± 0.002	3.8 ± 0.2	18.2 ± 0.5	473
DistilBERT	0.884 ± 0.006	0.875 ± 0.006	0.881 ± 0.006	0.878 ± 0.006	0.938 ± 0.003	2.1 ± 0.1	9.4 ± 0.3	265
MentalBERT	0.920 ± 0.005	0.913 ± 0.005	0.924 ± 0.005	0.918 ± 0.005	0.962 ± 0.002	4.0 ± 0.2	19.7 ± 0.6	475

False negatives often arise from indirect, understated, or context-dependent expressions of distress, including implicit hopelessness, short, fragmented statements, and multi-clause constructions where the depressive cue is subtle. Another common source of errors is negation and contrast (e.g., "I'm not sad anymore"), where surface-level sentiment may be misleading. In these scenarios, classical architectures such as CNN and LSTM more frequently fail to resolve long-range dependencies or pragmatic intent, whereas transformer-based models tend to be more resilient due to contextual encoding—although pragmatic ambiguity, figurative language, and complex discourse structure remain challenging even for contextual encoders.

Overall, the error analysis suggests that robust depression-related text classification benefits from strong contextual modeling, but additional improvements may require explicit handling of negation, pragmatics, and user-level temporal context, especially when transferring across platforms and writing styles.

H. Cross-Validation and Cross-Platform Evaluation of Model Robustness

To assess generalization and stability beyond a single dataset, we evaluated model robustness using two complementary settings: within-dataset cross-validation and cross-platform transfer. Within-dataset cross-validation estimates how stable model performance is when training and testing are performed on different folds of the same corpus. Cross-platform transfer evaluates domain shift by training on one platform-specific dataset and testing on another, which better reflects real-world variability in vocabulary, style, and topic distribution.

Cross-platform transfer results in Table V indicate a consistent performance drop for all models relative to in-domain evaluation, highlighting a non-trivial domain shift between Reddit, Twitter, and Facebook. For within-domain validation, we conduct 5-fold stratified cross-validation on the merged Reddit+Twitter corpus and report the mean F1-score across folds. For cross-platform transfer, we (1) train on Reddit and evaluate on Twitter and (2) train on Twitter and evaluate on Reddit without target-domain fine-tuning. Additionally, to assess transfer to Facebook under a different discourse style, we train on the combined Reddit+Twitter corpus and evaluate on Facebook without collecting new Facebook data. Table V reports F1-scores for each setting.

TABLE V. CROSS-VALIDATION AND CROSS-PLATFORM PERFORMANCE OF NEURAL ARCHITECTURES (MEAN F1-SCORE, %)

Model	5-Fold CV (Reddit+Twitter) F1 (%)	Reddit→Twitter F1 (%)	Twitter→Reddit F1 (%)	Reddit+Twitter→Facebook F1 (%)
CNN	82.8	79.3	80.1	78.6
LSTM	86.7	82.6	83.4	82.0
BERT	92.4	86.8	87.5	85.9
RoBERTa	93.3	88.1	88.9	87.4
DistilBERT	90.4	84.9	85.7	84.1
MentalBERT	**94.0**	**89.3**	**90.1**	**88.8**

I. Ethical Considerations and Data Protection

This study uses research corpora described in prior work and accessed under their respective terms (public releases where available, or research-accessible benchmarks subject to platform and sharing constraints). We follow established internet-research ethics guidance and health-focused social-media research protocols that emphasize contextual integrity, risk assessment, and harm minimization when working with online traces in sensitive domains such as mental health. Our analysis is conducted at an aggregated level for methodological research rather than individual-level profiling or intervention. Where present, we remove or mask explicit identifiers (e.g., user handles, mentions, and URLs) during local processing, and we avoid reporting verbatim user-generated text; when illustrative examples are necessary, they are paraphrased to reduce re-identification risk and unintended disclosure [38], [39], [40].

Because mental-health–related language can be sensitive, we emphasize that model outputs should not be interpreted as clinical diagnoses. Potential harms include false positives and false negatives, which may lead to inappropriate concern or missed cases if deployed without safeguards. Therefore, any practical use should include human oversight, careful threshold calibration, and evaluation on the target domain and language before deployment [40]. We also recommend additional validation across platforms and demographic contexts to reduce bias and improve robustness.

We acknowledge that language models may inherit demographic, cultural, or platform-specific biases present in the underlying corpora. Therefore, fairness-sensitive evaluation and cross-domain validation are necessary before operational use. Any deployment in mental-health–related settings should incorporate human oversight, transparent reporting, and risk-mitigation procedures appropriate for sensitive decisions.

V. DISCUSSION

A. Main Findings and Interpretation

This study compared baseline neural architectures (CNN, LSTM, and BERT) [35] and several transformer variants (RoBERTa, DistilBERT, and MentalBERT) for classifying depression-related language in social-media text. Across the baseline comparison (Table III), BERT achieved the strongest overall balance in predictive quality, particularly in F1-score and AUC. This result is consistent with the advantage of contextual self-attention, which supports modeling non-local dependencies

and subtle semantic cues that frequently appear in user-generated mental-health narratives.

CNN showed the weakest performance among the three baselines. While convolutional filters can capture discriminative local patterns, the model is limited in representing long-range dependencies and context interactions that are important for distinguishing transient negative sentiment from depression-related expression. LSTM provided an intermediate solution: its recurrent structure better captures sequential context than CNN, improving Recall and AUC, yet it remained below transformer-based encoders in the main metrics. Importantly, these findings reinforce the practical view that model choice should be aligned with the intended deployment constraints: CNN and LSTM remain attractive when compute and latency budgets are tight, whereas transformers are preferable when higher screening sensitivity and overall separability are required.

The robustness-oriented experiments further clarify trade-offs among transformer variants (Tables III–IV). RoBERTa showed strong predictive performance as a general-purpose transformer, while DistilBERT provided a favorable efficiency profile with reduced inference latency and smaller model size at a moderate performance cost. MentalBERT achieved the highest quality among the tested variants in our setting, which is consistent with the expected benefit of domain-adaptive pretraining for mental-health language, where indirect phrasing, self-referential statements, and psychological terminology can be informative. At the same time, performance differences across variants should be interpreted together with computational indicators: in deployment-oriented scenarios, the "best" model depends on whether the priority is maximum predictive quality, minimal latency, or memory footprint. A plausible explanation is that domain-adaptive pretraining exposes the encoder to mental-health–specific vocabulary, self-referential patterns, and community discourse that are underrepresented in general-domain corpora. This can improve sensitivity to subtle distress cues that are difficult to capture with shallow n-gram features or purely sequential baselines.

The cross-validation and cross-platform results (Table V) highlight the impact of domain shift. Performance typically decreases when transferring across platforms, reflecting differences in writing style, slang, message length, and topic distribution. For example, for BERT, the within-domain cross-validation F1 is 92.4%, while transfer yields 86.8% (Reddit→Twitter) and 87.5% (Twitter→Reddit), corresponding to drops of 4.9 – 5.6 F1 points; transfer to Facebook is 85.9% (drop 6.5 points). The transformer family generally exhibited stronger robustness under transfer than CNN/LSTM, suggesting that contextual representations can reduce sensitivity to lexical variability. However, non-trivial domain shift remains a challenge, and platform-specific adaptation, threshold calibration, or additional domain-relevant pretraining may be necessary for stable real-world performance. These results suggest that strong in-domain performance alone is insufficient for deployment-oriented use cases and that robustness under domain shift should be treated as a first-class evaluation criterion.

B. Limitations

Several limitations should be noted. First, our experiments use post-level samples derived from user-level public corpora, and specific sampling, filtering, and deduplication decisions may affect absolute performance values across platforms. Second, labels in social-media datasets represent operational definitions from prior studies and are imperfect proxies for clinical diagnosis; they may contain label noise and self-report bias, and model outputs should be interpreted as research-oriented risk signals rather than diagnostic conclusions. Third, class imbalance in real deployments may differ from both the merged corpus and our fixed evaluation split; while we applied balancing only on the training split, we did not conduct a full probability calibration study (e.g., reliability curves or ECE), which is important for threshold-based triage settings. Fourth, CNN and LSTM baselines were implemented as standard lightweight architectures and are not strictly capacity-matched to transformer encoders; stronger non-transformer baselines (deeper CNNs, larger BiLSTMs, or modern efficient transformers) are a useful extension. Fifth, platform policies and data-sharing constraints limit redistribution of raw text and may restrict independent instance-level verification; accordingly, we report results at an aggregated level. Sixth, the unified corpus is English-only, and cross-lingual generalization remains an important direction for future work. Finally, although repeated-run reporting improves robustness, the number of runs ($n = 5$) is modest; future studies could extend variability analysis across additional seeds, report calibration under prevalence shift, and explore efficient deployment techniques such as quantization, pruning, and distillation under realistic batching and hardware constraints.

VI. Conclusion

This study presented a controlled comparison of CNN, LSTM, and transformer encoders for depression-related text classification on a unified multi-platform social-media corpus. Using a single-run baseline together with repeated-run evaluation and statistical testing, we found that transformer models, especially a domain-adapted encoder, deliver the strongest predictive performance and more stable behavior under cross-platform transfer. At the same time, transfer experiments showed a consistent degradation across platforms, indicating that deployment-oriented evaluation must go beyond in-domain accuracy and explicitly account for domain shift and robustness. Finally, our efficiency analysis highlights practical trade-offs between quality and computational cost, enabling informed model choice for real-world research pipelines. Future work will extend evaluation to additional languages and platforms, incorporate longitudinal user-level signals, and study calibration and fairness-sensitive safeguards for responsible use.

Acknowledgment

The authors thank colleagues and collaborators for helpful feedback on earlier drafts and the research community for releasing public resources that enable reproducible computational mental-health research. No external funding was received for this study.

REFERENCES

[1] World Health Organization, "Depressive disorder (depression)," Fact sheet, Aug. 29, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression (accessed Jan. 2026).

[2] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," Proc. Int. AAAI Conf. Web and Social Media (ICWSM), pp. 128–137, 2013, doi: 10.1609/icwsm.v7i1.14432.

[3] A.-M. Bucur, I. R. Podină, and L. P. Dinu, "A psychologically informed part-of-speech analysis of depression in social media," arXiv:2108.00279, 2021, doi: 10.48550/arXiv.2108.00279.

[4] G. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," npj Digital Medicine, vol. 3, no. 1, pp. 1–11, 2020, doi: 10.1038/s41746-020-0253-0.

[5] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, and P. S. Yu, "A survey on text classification: From shallow to deep learning," arXiv:2008.00364, 2020, doi: 10.48550/arXiv.2008.00364.

[6] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–based Text Classification: A Comprehensive Review," ACM Computing Surveys, vol. 54, no. 3, pp. 1–40, 2021, doi: 10.1145/3439726.

[7] A. Yates, A. Cohan, and N. Goharian, "Depression and Self-Harm Risk Assessment in Online Forums," in Proc. EMNLP 2017, pp. 2968–2978, 2017, doi: 10.18653/v1/D17-1321.

[8] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP 2014, pp. 1746–1751, 2014.

[9] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.

[10] I. Tavchioski, M. Robnik-Šikonja, and S. Pollak, "Detection of depression on social networks using transformers and ensembles," arXiv:2305.05325, 2023, doi: 10.48550/arXiv.2305.05325

[11] A. Vaswani et al., "Attention Is All You Need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 5998–6008, 2017, doi: 10.48550/arXiv.1706.03762.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT 2019, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[13] S. Ji et al., "MentalBERT: Publicly available pretrained language models for mental healthcare," arXiv:2110.15621, 2021, doi: 10.48550/arXiv.2110.15621.

[14] S. Munir, S. A. Gillani, M. S. A. Baig, M. A. Saleem, and H. Siddiqui, "Advancing Depression Detection on Social Media Platforms through Fine-Tuned Large Language Models," arXiv:2409.14794, 2024, doi: 10.48550/arXiv.2409.14794.

[15] R. W. Leiva, A. Martín, and A. González, "Deep learning for depression detection from text: A comparative analysis," Expert Systems with Applications, vol. 183, p. 115376, 2021, doi: 10.1016/j.eswa.2021.115376.

[16] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," in Proc. ACL 2018, pp. 1383–1392, doi: 10.18653/v1/P18-1128.

[17] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey," ACM Computing Surveys, vol. 55, no. 6, Art. no. 109, 2022, doi: 10.1145/3530811.

[18] S. C. Guntuku, D. B. Yaden, M. Kern, L. H. Ungar, and H. A. Schwartz, "Detecting depression and mental illness on social media: An integrative review," Current Opinion in Behavioral Sciences, vol. 18, pp. 43–49, 2017, doi: 10.1016/j.cobeha.2017.07.005.

[19] Suhavi et al., "Twitter-STMHD: An extensive user-level database of multiple mental health disorders," in Proc. ICWSM, vol. 16, pp. 1182–1191, 2022. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19368

[20] 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 Proc. LREC-COLING 2024, Torino, Italy, May 2024, pp. 11196–11203.

[21] E. Bao, A. Pérez, and J. Parapar, "ReDSM5: A Reddit Dataset for DSM-5 Depression Detection," arXiv:2508.03399, Aug. 2025. [Online]. Available: https://arxiv.org/abs/2508.03399

[22] B. Wang, Y. Sun, Y. Zhao, and B. Qin, "Beyond Snapshots: A Multimodal User-Level Dataset for Depression Detection in Dynamic Social Media Streams," in Proc. ACM Multimedia (MM '25), Oct. 2025, doi: 10.1145/3746027.3758236.

[23] T. Tseriotou et al., "Overview of the CLPsych 2025 Shared Task: Capturing Mental Health Dynamics from Social Media Timelines," in Proc. CLPsych 2025, May 2025, pp. 193–217.

[24] S. Mankarious and A. Zirikly, "CARMA: Comprehensive Automatically-annotated Reddit Mental Health Dataset for Arabic," arXiv:2511.03102, Nov. 2025. [Online]. Available: https://arxiv.org/abs/2511.03102

[25] "eRisk: Early Risk Prediction on the Internet – CLEF 2025," CLEF 2025 Labs. [Online]. Available: https://clef2025.clef-initiative.eu/index.php?page=Pages/Labs/eRisk.html (accessed Dec. 16, 2025).

[26] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," Health Information Science and Systems, vol. 6, no. 1, Art. no. 8, 2018, doi: 10.1007/s13755-018-0046-0.

[27] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning from Imbalanced Data Sets. Springer, Cham, 2018, doi: 10.1007/978-3-319-98074-4.

[28] S. Y. Feng et al., "A Survey of Data Augmentation Approaches for NLP," in Findings of ACL-IJCNLP 2021, pp. 968–988, 2021, doi: 10.18653/v1/2021.findings-acl.84.

[29] E. Ma, "nlpaug: NLP augmentation library," GitHub repository, 2019. [Online]. Available: https://github.com/makcedward/nlpaug (accessed Dec. 2025).

[30] H. E. Tay, M. K. Lim, and C. Y. Chong, "SERCNN: Stacked Embedding Recurrent Convolutional Neural Network in Detecting Depression on Twitter," arXiv:2207.14535, 2022, doi: 10.48550/arXiv.2207.14535.

[31] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 1–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[32] N. Reimers and I. Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging," in Proc. EMNLP 2017, pp. 338–348, 2017, doi: 10.18653/v1/D17-1035.

[33] J. Dodge et al., "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping," arXiv:2002.06305, 2020. [Online]. Available: https://arxiv.org/abs/2002.06305

[34] M. Hollander, D. A. Wolfe, and E. Chicken, Nonparametric Statistical Methods, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[35] R. Dror, S. Shlomov, and R. Reichart, "Deep Dominance - How to Properly Compare Deep Neural Models," in Proc. ACL 2019, pp. 2773–2785, 2019, doi: 10.18653/v1/P19-1266.

[36] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[38] A. S. Franzke, A. Bechmann, C. M. Ess, and M. Zimmer (Eds.), Internet Research: Ethical Guidelines 3.0. Association of Internet Researchers (AoIR), 2019.

[39] L. Townsend and C. Wallace, Social Media Research: A Guide to Ethics. University of Aberdeen, 2016.

[40] A. Benton, G. Coppersmith, and M. Dredze, "Ethical Research Protocols for Social Media Health Research," in Proc. EthNLP (ACL Workshop), Valencia, Spain, 2017, pp. 94–102, doi: 10.18653/v1/W17-1612.