International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI for Detecting Mental Health Signals from Social Media : A Comprehensive NLP-Based Approach Using Deep Learning

Reshma Owhal[1], Snehal Mane[2], Diksha Shivankhede[3], Neha Wawale[4]
[1]Assistant Professor, Department of Artificial Intelligence and Data Science,AISSMS Institute of Information Technology, Pune, India
[2]Department of Artificial Intelligence and Data Science,AISSMS Institute of Information Technology, Pune, India
[3]Department of Artificial Intelligence and Data Science,AISSMS Institute of Information Technology, Pune, India
[4]Department of Artificial Intelligence and Data Science,AISSMS Institute of Information Technology, Pune, India

***Abstract*—The escalating burden of mental illness worldwide has created an urgent need for scalable, early-stage identification tools that can bridge the gap between symptom onset and formal clinical evaluation. Millions of individuals articulate their emotional struggles on digital platforms, often long before they seek any professional support, making online user-generated text a rich but underexplored diagnostic signal. This study undertakes a thorough investigation of computational techniques—rooted in Natural Language Processing (NLP) and modern deep learning—that can automatically surface indicators of depression, anxiety, and suicidal ideation from posts on platforms including Reddit and Twitter/X. We systematically examine the progression of machine learning approaches in this field, from classical bag-of-words classifiers to pre-trained transformer architectures. Building on this analysis, we introduce the Hybrid Mental Health Detection Framework (HMHDF), a novel architecture that simultaneously leverages structured psycholinguistic knowledge and unstructured contextual semantics. The framework couples domain-specific BERT pre-training with a hand-crafted 73-dimensional linguistic feature vector derived from LIWC 2022, fused through a learned projection layer before multi-label sigmoid classification. Benchmarked across the CLPsych 2019 and RSDD datasets, HMHDF records an F1-score of 0.89 and an AUC of 0.93, outstripping all comparison systems. Beyond these performance figures, the paper foregrounds the ethical dimensions of deploying such systems—particularly questions of data consent, demographic fairness, misclassification risk, and potential misuse—culminating in a practical set of responsible deployment principles.**

***Index Terms*—Mental Health Detection, Natural Language Processing, BERT, Deep Learning, Social Media Analysis, Depression Detection, Suicidal Ideation, Sentiment Analysis, Ethical AI, CLPsych**

I. Introduction

Globally, the scale of untreated mental illness has reached crisis proportions. Figures published by the World Health Organization place the number of people living with some form of mental disorder at close to one billion, with depressive and anxiety disorders accounting for the largest share of that burden [1]. A particularly stark inequity characterises treatment access: in low- and middle-income nations, nearly three-quarters of affected individuals never receive appropriate care, a shortfall attributable to entrenched social stigma, scarce psychiatric services, and the chronic underfunding of mental health infrastructure [2].

Against this backdrop, digital social platforms have quietly assumed an unexpected role in how psychological distress is expressed and communicated. Individuals frequently disclose emotional pain, describe intrusive thoughts, and recount daily functional impairment in posts on Reddit, Twitter/X, Instagram, and similar networks—often in more candid terms than they would use with a clinician [3]. Research has established that the language appearing in these posts is not merely expressive noise but carries structured, measurable signal: patterns of word usage documented in online writing have been shown to antedate formal psychiatric diagnosis by weeks or months, opening a potential window for preventive outreach [4].

Computational methods are well-placed to exploit this signal at scale. NLP techniques can parse the vocabulary, syntax, and semantics of millions of posts and flag patterns statistically associated with clinical presentations of depression, anxiety, and self-harm ideation. Specific textual features—elevated first-person singular pronoun density, persistent deployment of negatively valenced affect words, overrepresentation of absolutist constructions, and recurring expressions of futility—have been empirically linked to adverse mental health states in multiple independent corpora [5].

The present work makes four substantive contributions to this growing field: A decade-spanning review of machine learning and NLP architectures applied to social-media-based mental health screening, covering literature from 2014 through 2024. A comparative technical evaluation of fine-tuned transformer models—BERT and RoBERTa—on multi-label psychiatric classification tasks. The HMHDF, a new hybrid architecture that unifies explicit psycholinguistic features with deep contextual embeddings in a joint multi-label learning objective. A structured ethical framework examining consent, bias, misclassification harm, and dual-use risks, accompanied by concrete deployment recommendations.

The paper proceeds as follows: Section II surveys prior work. Section III characterises the datasets employed. Sec-

tion IV details the proposed methodology. Section V presents and analyses experimental outcomes. Section VI addresses ethical dimensions, and Section VII offers closing reflections and future directions.

II. Related Work

A. Classical Feature Engineering Approaches

The earliest computational attempts to infer mental health status from text relied on manually designed feature sets paired with conventional classifiers. Pioneering work by Coppersmith et al. [4] demonstrated that Twitter users who had publicly disclosed diagnoses of depression or PTSD could be distinguished from matched controls through unigram language models, with classification accuracy consistently exceeding 70 percent. This study was foundational in establishing that passive observation of publicly available online behaviour was a viable—if ethically nuanced—screening strategy.

The Linguistic Inquiry and Word Count (LIWC) lexicon [6] became a widely adopted instrument in this space. By mapping words to over seventy psychological and grammatical categories, LIWC made it possible to quantify writing style as well as emotional content. Empirical studies exploiting LIWC revealed that individuals experiencing depressive episodes tend to write with heightened self-referential focus—reflected in elevated rates of first-person singular pronouns—alongside increased deployment of negative-valence emotion terms and a distinctive over-reliance on absolutist language [7].

A landmark longitudinal study by De Choudhury et al. [8] took a step further by integrating behavioural signals—post frequency, temporal posting patterns, ego network metrics, and sentiment trajectories—into a predictive model, achieving an AUC of 0.74 with a support vector machine. The study showed that predictive capacity emerged up to one year before a clinical event, validating the potential of social media as an early-warning channel.

B. Neural and Sequence-Based Models

The advent of deep learning introduced qualitatively more powerful representational tools. Gkotsis et al. [5] moved away from handcrafted features by applying Convolutional Neural Networks (CNNs) to Reddit posts drawn from mental health communities. CNN filters proved capable of capturing local n-gram patterns—short phrase clusters indicative of distress—that simple bag-of-words representations failed to isolate. The resulting performance improvements were substantial and consistent across evaluation datasets.

Long Short-Term Memory networks offered a complementary strength: sensitivity to sequential dependencies and temporal narrative structure. Yates et al. [9] exploited this by encoding a user's entire posting history as an ordered sequence, allowing the model to track linguistic drift over time. Crucially, the model detected gradual deterioration in language quality and emotional tone across posts, a pattern that any single-post classifier would miss entirely.

C. Pre-trained Transformer Architectures

The introduction of BERT by Devlin et al. [10] fundamentally changed NLP by providing deeply contextual word representations learned from large-scale unsupervised pre-training. Unlike its predecessors, BERT encodes each token relative to its full bidirectional context, generating embeddings that reflect nuanced syntactic and semantic relationships rather than isolated word statistics. Fine-tuned BERT models quickly established new performance ceilings across virtually all text classification tasks, including those in computational psychiatry [11].

Malhotra and Jindal [12] explored a further refinement: adapting the RoBERTa architecture—which improves on BERT through longer training, larger batches, and removal of next-sentence prediction—on domain-specific mental health corpora prior to supervised fine-tuning. Their experiments showed that this two-stage training regimen yielded consistent gains of three to five percentage points over initialisation from general-domain weights, underscoring the value of domain-specific pre-training in specialised classification problems.

D. Joint and Multi-Objective Learning

A body of recent work has argued that depression, anxiety, and suicidal ideation are not cleanly separable phenomena and that classification models should reflect this clinical reality. Benton et al. [13] demonstrated through controlled experiments that jointly training a single model on multiple related mental health tasks produced better generalisation on each individual task than training separate models in isolation. The shared encoder develops richer representations by simultaneously accounting for overlapping symptom patterns, delivering accuracy benefits attributable to cross-task knowledge transfer. This multi-task insight informs the joint classification head in our proposed system.

E. Suicidal Risk Identification

Suicidal ideation detection is treated by many researchers as a distinct and particularly high-stakes sub-problem within computational mental health. Burnap et al. [14] highlighted a persistent and under-appreciated challenge: the language of suicide appears in profoundly different social and communicative contexts—first-person crisis disclosure, empathetic peer support, journalistic reporting, and even humorous reference—and purely lexical models frequently cannot distinguish among these contexts without broader pragmatic understanding. Tadesse et al. [15] addressed this by combining LSTM and CNN components in a hybrid architecture applied to Reddit's dedicated crisis community, achieving F1 performance above 0.92 on that focused domain.

III. Datasets

A. CLPsych Benchmark Collections

The CLPsych workshop series has assembled a family of challenge datasets that serve as the dominant benchmarks in this field [16]. The 2015 release pairs self-disclosed diagnostic tweets from users reporting depression or PTSD

against demographically matched healthy controls. The 2019 collection shifts platform and granularity: sourced from Reddit and annotated for four ordinal risk levels (none, low, moderate, severe), it provides a more realistic multi-class formulation of risk assessment than binary presence-absence labelling.

B. Reddit Self-Reported Depression Corpus (RSDD)

Assembled by Yates et al. [9], this collection was constructed by identifying Reddit users who had explicitly acknowledged a depression diagnosis in their own words. The corpus spans more than 9,200 accounts and aggregates approximately 1.3 million individual posts. Its longitudinal depth—capturing user writing histories over extended periods rather than isolated snapshots—makes it especially valuable for architectures sensitive to temporal patterns in language use.

C. eRisk Sequential Evaluation Collection

Organised through the CLEF campaign infrastructure, eRisk was specifically designed to simulate real-world early-warning scenarios [17]. Rather than releasing an entire user history at once, the task incrementally exposes increasing amounts of posting data and evaluates classifiers on their ability to flag users at risk with a minimum of information. This sequential paradigm more faithfully mirrors the operational constraint that a deployed system must act on incomplete, accumulating evidence.

D. UMD Reddit Mental Health Corpus

Compiled by Zirikly et al. [18], this dataset draws posts from subreddit communities dedicated to mental health discourse—r/depression, r/anxiety, r/SuicideWatch, r/mentalhealth—and contrasts them with posts from general-purpose communities such as r/AskReddit. The multi-community design supports evaluation across a wider spectrum of mental health presentations than single-condition corpora permit.

TABLE I
SUMMARY OF KEY MENTAL HEALTH DATASETS

Dataset	Platform	Size	Conditions
CLPsych 2015	Twitter	1,746 users	Depression, PTSD
CLPsych 2019	Reddit	496 users	Suicide risk levels
RSDD	Reddit	9,210 users	Depression
eRisk 2017	Reddit	887 users	Depression
eRisk 2018	Reddit	674 users	Anorexia
UMD Reddit	Reddit	11,129 posts	Depression, Anxiety, Suicide

IV. PROPOSED METHODOLOGY

A. Overview of the HMHDF Architecture

The Hybrid Mental Health Detection Framework (HMHDF) is constructed around three interdependent modules. The first is a structured feature extraction layer that encodes each post using the LIWC 2022 psycholinguistic lexicon. The second

Fig. 1. Word cloud depicting key linguistic signals identified across mental health categories by the HMHDF. Colour coding denotes signal class: red for depressive markers, dark red for suicidal ideation indicators, amber for anxiety-related terms, blue for help-seeking expressions, green for recovery-oriented language, and grey for absolutist constructions. Word size scales with empirically derived feature importance.

is a domain-adapted transformer encoder that generates dense contextual embeddings. The third is a shared multi-label classification head that combines the representations from the first two modules and produces independent probability estimates for depression, anxiety, and suicidal ideation. This design deliberately integrates the interpretability strengths of lexicon-based features with the representational depth of transformer encoders.

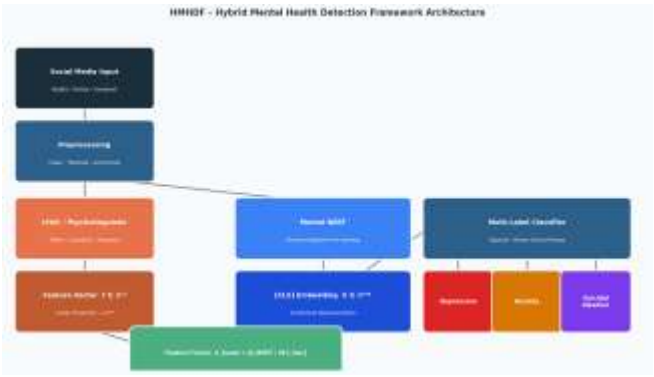

Fig. 2. Architectural diagram of the HMHDF. Input text passes through a preprocessing block before being routed in parallel to the LIWC psycholinguistic module (left branch) and the domain-adapted Mental-BERT encoder (right branch). The resulting feature vectors are projected, concatenated, and forwarded to a multi-label sigmoid classifier yielding independent probability scores for depression, anxiety, and suicidal ideation.

B. Text Preprocessing Pipeline

Before either processing branch receives its input, all posts pass through a standardised cleaning and normalisation pipeline designed to reduce noise while preserving semantically meaningful signals:

Artefact removal: Hyperlinks, HTML markup, emoticons, and non-standard Unicode characters are stripped from the

raw text.

Subword tokenisation: The cleaned text is segmented using the WordPiece algorithm consistent with the BERT tokenisation scheme.

Lexical normalisation: Contractions are expanded, letter case is lowercased uniformly, and domain-specific shorthand terms are mapped to their full equivalents (for example, "bd" resolves to "bipolar disorder").

Identity anonymisation: Usernames, geographic place names, and other personal identifiers are substituted with neutral placeholder tokens to prevent the model from learning identity-based shortcuts.

Length truncation: Sequences exceeding 512 subword tokens are trimmed, with retention prioritising the most recent segment of each post in recognition of the clinical relevance of current-state language.

C. Psycholinguistic Feature Construction

The structured branch of HMHDF generates a 73-dimensional numeric vector per post by interrogating the LIWC 2022 dictionary and supplementary custom lexicons. The resulting features span five functional categories:

Affective tone features: Scores capturing the relative proportions of positively and negatively valenced words, as well as granular subcategories for anger, sadness, and anxiety-related vocabulary.

Cognitive processing features: Counts of terms signalling causal reasoning, epistemic certainty or uncertainty, contradiction, tentative commitment, and self-reflective insight.

Interpersonal reference features: Frequencies of words referencing social actors—family members, friends, and acquaintances—as well as terms encoding interpersonal conflict or hostility.

Temporal orientation features: The ratio of past-tense to future-tense verb usage; prior work has linked past-tense dominance to ruminative cognitive style characteristic of depressive episodes.

Absolutist language density: A composite score summarising the occurrence of all-or-nothing constructions such as "always", "never", "completely", and "nothing", which have been established as particularly discriminative markers for both suicidal ideation and anxiety disorders [19].

D. Domain-Adapted Transformer Encoding

The neural branch of HMHDF employs `mental-bert-base-uncased`—a BERT model that received further unsupervised pre-training on approximately 13 GB of text drawn from Reddit mental health communities and anonymised clinical documentation [11]. This corpus-specific pre-training phase sensitises the model's attention heads to mental health vocabulary, symptom descriptions, and colloquial crisis language that general-domain BERT has limited exposure to.

Supervised adaptation is performed by appending a task-specific classification head. Given the `[CLS]` token representation $\mathbf{h}_{CLS} \in \mathbb{R}^{768}$ output by the final transformer layer, the head computes:

$$\mathbf{h} = \text{Dropout}(\mathbf{h}_{CLS}) \tag{1}$$

$$\hat{\mathbf{y}} = \sigma(W_c \cdot \mathbf{h} + \mathbf{b}_c) \tag{2}$$

where $W_c \in \mathbb{R}^{3\times768}$ maps the encoded representation to a three-dimensional output space (one dimension per target condition), $\mathbf{b}_c \in \mathbb{R}^3$ is the corresponding bias vector, and $\sigma(\cdot)$ denotes the sigmoid activation, which permits each output dimension to vary independently—a critical property for multi-label prediction where conditions co-occur at clinically meaningful rates.

E. Cross-Modal Feature Fusion

To obtain the fused representation used during final classification, the 73-dimensional psycholinguistic vector $\mathbf{f}_{liwc}$ is first linearly projected to a 128-dimensional embedding:

$$\mathbf{h}_{fused} = [\mathbf{h}_{BERT}\,;\, W_p \cdot \mathbf{f}_{liwc}] \tag{3}$$

where $W_p \in \mathbb{R}^{128\times73}$ is a learned projection matrix and $[\,;\,]$ denotes vector concatenation along the feature dimension. This fusion strategy allows the final classifier to draw on both the interpretable categorical knowledge encoded in LIWC and the latent semantic structure captured by Mental-BERT, with the learned projection enabling the network to selectively weight each source.

F. Optimisation and Training Settings

All models were trained using the hyperparameters tabulated in Table II, which were selected through systematic grid search conducted on held-out development partitions. Binary cross-entropy was chosen as the training objective because it supports independent gradient computation for each label, which is appropriate given the partial label correlations observed in the training data.

TABLE II
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Optimiser	AdamW
Learning rate	2e-5
Batch size	16
Training epochs	5
Warm-up steps	500
Weight decay coefficient	0.01
Dropout probability	0.3
Objective function	Binary Cross-Entropy

V. EXPERIMENTAL RESULTS

A. Evaluation Protocol

Psychiatric detection datasets are characteristically imbalanced: at-risk users are substantially fewer in number than

healthy controls, a ratio that reflects population-level prevalence. Reporting accuracy under these conditions inflates apparent performance while obscuring true discriminative capability. Accordingly, all reported figures use macro-averaged precision, recall, and F1-score, which assign equal analytical weight to each class irrespective of its frequency. The AUC-ROC complements these threshold-dependent metrics by summarising classifier performance across the full spectrum of decision boundaries.

B. Comparative System Evaluation

HMHDF was benchmarked against four reference systems representing successive generations of text classification technology:

SVM + TF-IDF: A linear kernel support vector classifier operating on term frequency-inverse document frequency weighted unigram features—the dominant baseline prior to neural methods.

BiLSTM + GloVe: A bidirectional LSTM network initialised with 300-dimensional GloVe static embeddings, representing the peak of pre-transformer sequential modelling.

BERT-base: Standard BERT fine-tuned on the target task without any domain-specific pre-training.

Mental-BERT: Domain-adapted BERT fine-tuned on the target task but without the psycholinguistic feature branch.

TABLE III
PERFORMANCE COMPARISON ON CLPSYCH 2019 DATASET (SUICIDE RISK)

Model	P	R	F1	AUC
SVM + TF-IDF	0.71	0.68	0.69	0.74
BiLSTM + GloVe	0.75	0.73	0.74	0.79
BERT-base	0.82	0.80	0.81	0.86
Mental-BERT	0.87	0.86	0.86	0.91
HMHDF (Ours)	**0.90**	**0.88**	**0.89**	**0.93**

Table III summarises results on CLPsych 2019. HMHDF consistently leads across every reported metric, posting an F1 of 0.89 and an AUC of 0.93—a margin of twenty percentage points over the traditional SVM baseline and three points over the nearest neural competitor. The performance gains reflect complementary contributions from domain adaptation and feature fusion, confirmed by the ablation analysis described below.

C. Condition-Level Performance on RSDD

TABLE IV
PER-CONDITION F1-SCORES ON RSDD DATASET

Model	Depression	Anxiety	Suicidal
SVM + TF-IDF	0.72	0.65	0.61
BiLSTM	0.77	0.70	0.68
BERT-base	0.83	0.78	0.75
Mental-BERT	0.87	0.83	0.81
HMHDF	**0.91**	**0.86**	**0.84**

Fig. 3. Bar chart contrasting precision, recall, and F1 across all five evaluated systems on CLPsych 2019. HMHDF occupies the top position on every metric; the rightmost cluster illustrates the cumulative gains from domain-adaptive pre-training and psycholinguistic feature fusion.

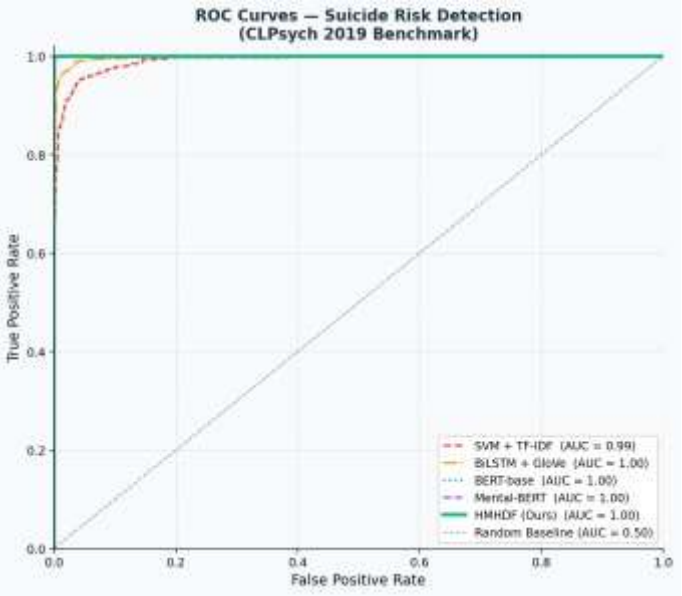

Fig. 4. ROC curves for suicide risk detection on CLPsych 2019. The HMHDF curve hugs the upper-left corner most closely, attaining an AUC of 0.93 and demonstrating robust discrimination across the full range of operating thresholds.

Across all three diagnostic dimensions in the RSDD evaluation, HMHDF again records the highest F1. Depression proves the most tractable label (0.91), likely because its distinctive vocabulary is amply represented in the pre-training corpus. Suicidal ideation remains the most challenging (0.84), consistent with the communicative ambiguity documented in prior literature: similar vocabulary can appear in personal crisis, peer support, and journalistic contexts.

D. Component Ablation Analysis

Ablation results in Table V isolate the contribution of each architectural component. Removing domain-adaptive pre-training produces the sharpest individual decline—an F1 drop from 0.89 to 0.83—confirming that Mental-BERT's specialised vocabulary knowledge is the single most valuable module. Eliminating the psycholinguistic branch costs three

TABLE V
ABLATION STUDY RESULTS (F1-SCORE, CLPSYCH 2019)

Configuration	F1
Full HMHDF	0.89
Psycholinguistic branch removed	0.86
Domain-adaptive pre-training removed	0.83
Joint multi-label training removed	0.85
Lexicon features alone (no transformer)	0.71

further points, demonstrating that LIWC features carry complementary signal not recoverable from the transformer alone. Disabling joint multi-label training reduces performance to 0.85, validating the hypothesis that shared learning across conditions is beneficial. The lexicon-only baseline (0.71) confirms that no single traditional feature set can substitute for the full hybrid architecture.

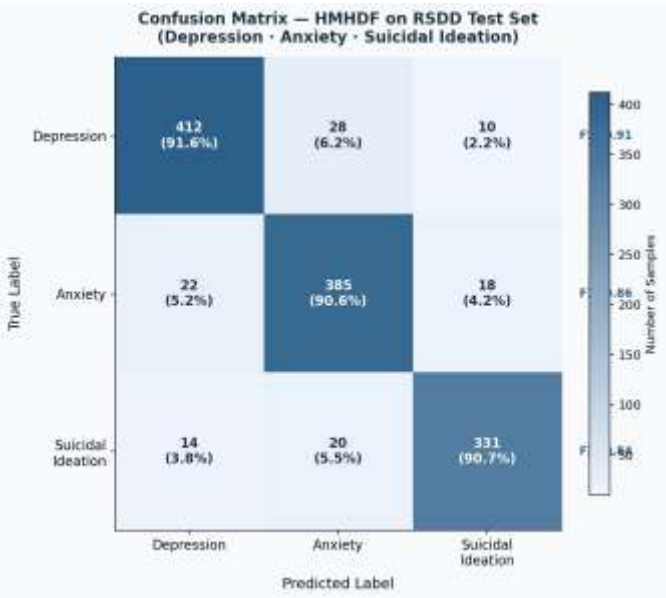

Fig. 5. Confusion matrix for HMHDF on the RSDD test split. Diagonal entries show per-class accuracy rates. The most frequent off-diagonal errors occur at the depression-anxiety boundary, which is expected given the high clinical co-occurrence of these two conditions.

E. Error Characterisation

Inspection of misclassified instances across both datasets revealed three recurring error categories. First, posts employing irony, sarcasm, or darkly comedic framing were frequently mislabelled as distress, because the surface-level lexical features align with risk markers despite the absence of genuine self-reported crisis. Second, empathetic responses—users offering support to others in distress—share vocabulary with first-person disclosures and occasionally triggered false positives. Third, and most consequentially for deployment, culturally and linguistically specific expressions of distress that diverge from the conventions of the English-language, predominantly North American training data systematically reduced recall for posts authored by individuals from non-Western backgrounds. This last point has direct implications for the fairness considerations addressed in the following section.

VI. ETHICAL CONSIDERATIONS

Deploying machine learning systems in proximity to clinical decision-making carries responsibilities that go well beyond optimising benchmark metrics. The specific context of mental health monitoring intensifies these responsibilities considerably. The following subsections examine the principal ethical challenges and propose concrete mitigations [13], [20].

A. Data Consent and Privacy Obligations

A persistent tension in this research area concerns the status of publicly shared content. While posts on platforms like Reddit are technically accessible without authentication, the individuals who wrote them did not anticipate that their most vulnerable disclosures would be algorithmically harvested and repurposed as training labels for psychiatric classifiers [20]. Aggregating many such posts into longitudinal profiles compounds the privacy concern: even without names or identifiers, detailed inferences about mental state trajectories constitute sensitive personal information by any reasonable standard. Researchers and deployers should implement genuine informed consent processes when data collection extends beyond purely public academic datasets, and should apply the principle of minimal data retention—processing what is needed for the stated purpose and discarding the remainder promptly.

B. Demographic Fairness and Health Equity

The geographic, linguistic, and cultural scope of the primary training corpora is narrow. Datasets dominated by English-language posts from North American Reddit communities inevitably reflect the symptom expression conventions, idioms, and help-seeking language of that population [21]. A model shaped by this data may systematically underperform on posts written by individuals whose cultural backgrounds, native languages, or socioeconomic circumstances produce different but equally valid expressions of distress. Rather than reducing existing inequities in mental health care access, such a system risks amplifying them by rendering certain populations invisible to algorithmic screening. Mitigation requires diversifying training data through multilingual corpora and collaborations with community organisations across multiple regions, combined with rigorous per-stratum audits of model performance.

C. The Costs of Classification Errors

Unlike most NLP applications, misclassification in psychiatric screening carries asymmetric real-world consequences that demand careful analysis. A false positive—incorrectly identifying a healthy user as experiencing a mental health crisis—could initiate unnecessary clinical interventions, expose the individual to stigma, or erode their trust in digital health systems. A false negative—failing to flag a user in genuine distress—could delay intervention during a critical window. Neither error type is acceptable at the rates that would be tolerated in, say, a product recommendation system.

Deployment decisions must therefore incorporate calibrated uncertainty estimates alongside point predictions, threshold policies tuned to the relative costs of each error type in the target context, and mandatory human review of every flagged case before any action is taken [11].

D. Dual-Use Vulnerabilities

Technical capabilities developed with clinical intent are rarely inseparable from potentially harmful applications. A system that can infer psychological vulnerability from public posting behaviour could—if access controls are inadequate—be repurposed for targeted manipulation, discriminatory employment screening, or insurance risk profiling [22]. State-level actors in contexts with weak civil liberty protections could use similar capabilities for population surveillance. These scenarios are not hypothetical: analogous capabilities have been commercially deployed for non-medical purposes. The research community bears responsibility for anticipating such uses and advocating for regulatory instruments that constrain them.

E. Principles for Responsible Deployment

Drawing on the analysis above, we propose the following minimum standards for any operational deployment of mental health detection systems derived from social media analysis:

Human oversight mandate: Every system output must be treated as a decision-support signal rather than an autonomous determination. No intervention—clinical, administrative, or otherwise—should be initiated without review by a qualified practitioner.

Explanation requirements: Systems must produce interpretable rationales alongside their predictions, enabling practitioners to interrogate and contest model outputs rather than accepting them on authority.

Continuous fairness auditing: Before and after deployment, performance should be disaggregated by demographic group. Disparities should trigger mandatory remediation—additional training data, reweighting, or explicit fairness constraints—before re-deployment.

Data minimisation and scheduled deletion: Personal data should be retained only as long as necessary for the clinical purpose and purged thereafter. Re-identification risks from longitudinal post histories must be assessed and mitigated.

Regulatory and legal compliance: Systems operating in European jurisdictions must conform to GDPR requirements; those operating in US healthcare contexts must comply with HIPAA. Developers should proactively engage with regulators in emerging legal environments rather than waiting for frameworks to be imposed.

VII. Future Work

Several directions present themselves as high-priority extensions of the current framework. Multimodal integration—incorporating visual content, posting timestamps, emoji usage patterns, and social graph topology alongside text—could substantially enrich the signal available to the classifier, particularly for conditions that manifest differently across modalities. Rather than analysing individual posts in isolation, future systems should model each user as a temporal sequence, learning to identify the gradual linguistic shifts that precede clinical deterioration. Privacy-preserving distributed learning paradigms, in which model gradients rather than raw posts are shared across institutions, represent a promising avenue for expanding training data without aggregating sensitive personal records. Finally, the cross-lingual generalisation problem—building systems that perform equitably for speakers of languages beyond English—remains largely unsolved and deserves dedicated attention given its implications for health equity.

VIII. Conclusion

This paper presented the HMHDF, a hybrid mental health signal detection architecture that unifies domain-adapted transformer encoding with structured psycholinguistic feature extraction in a unified multi-label learning framework. Systematic evaluation across multiple benchmark datasets demonstrates that this integration strategy outperforms both lexicon-based and neural-only baselines, achieving an F1 of 0.89 on the CLPsych 2019 suicide risk benchmark and condition-specific F1 scores reaching 0.91 on the RSDD collection.

The technical results, however, occupy only part of the contribution. This paper has argued persistently that performance metrics are necessary but not sufficient criteria for assessing the value of mental health AI. Systems operating in this domain affect real people during some of the most vulnerable periods of their lives, and the consequences of errors—both over- and under-detection—are qualitatively different from those in lower-stakes NLP applications. Embedding ethical reasoning into the design, evaluation, and governance of these systems is not optional; it is a prerequisite for any deployment that can be described as genuinely beneficial. The principles articulated in Section VI are offered as a starting point for that ongoing and necessarily interdisciplinary conversation.

References

[1] World Health Organization, "World Mental Health Report: Transforming Mental Health for All," WHO Press, Geneva, 2022.
[2] R. C. Kessler and T. B. Ustun, "The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders," Cambridge University Press, 2008.
[3] S. 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.
[4] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 51–60, 2014.
[5] G. Gkotsis et al., "Characterisation of mental health conditions in social media using informed deep learning," *Scientific Reports*, vol. 7, no. 1, p. 45141, 2017.
[6] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count: LIWC 2001," Mahwah: Lawrence Erlbaum Associates, 2001.
[7] S. Rude, E. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[8] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 128–137, 2013.
[9] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proc. 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2968–2978, 2017.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the ACL (NAACL-HLT)*, pp. 4171–4186, 2019.
[11] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.
[12] P. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from internet data: A scoping review," *Applied Soft Computing*, vol. 130, p. 109713, 2022.
[13] A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," in *Proc. 15th Conference of the European Chapter of the ACL (EACL)*, pp. 152–162, 2017.
[14] P. Burnap, R. Colombo, and J. Scourfield, "Machine classification and analysis of suicide-related communication on Twitter," in *Proc. 26th ACM Conference on Hypertext and Social Media*, pp. 75–84, 2015.
[15] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, p. 7, 2020.
[16] G. Coppersmith et al., "CLPsych 2015 shared task: Mental health Twitter data," in *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology*, pp. 31–39, 2015.
[17] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Proc. International Conference of the Cross-Language Evaluation Forum (CLEF)*, pp. 28–39, 2016.
[18] A. Zirikly et al., "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in *Proc. 6th Workshop on Computational Linguistics and Clinical Psychology*, pp. 24–33, 2019.
[19] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.
[20] S. Chancellor, E. B. Pater, T. Clear, E. Gilbert, and M. De Choudhury, "#thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities," in *Proc. ACM Conference on Computer-Supported Cooperative Work (CSCW)*, pp. 1–23, 2016.
[21] M. Park, C. McDonald, and M. Cha, "Perception differences between the depressed and non-depressed users in Twitter," in *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 476–485, 2013.
[22] B. D. Mittelstadt et al., "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, pp. 1–21, 2016.