KMAN PUBLISHING INSTITUTE
Journal Website

Article history:
Received 02 May 2026
Revised 03 August 2026
Accepted 10 August 2026
Initial Publication 15 August 2026
Final Publication 20 August 2026

KMAN Counseling & Psychology Nexus

Volume 4, pp 1-18

E-ISSN: 3041-9026

Explainable Semi-Supervised Learning for Depression Subtype Detection in Social Media

Ali. Nakhaei[1], Hamed. Vahdat Nejad[2*], Mahdi. Khazaei Pour[1]

[1] Department of Computer Engineering, Bi.C., Islamic Azad University, Birjand, Iran
[2] Department of Computer Engineering, University of Birjand, Birjand, Iran

*** Corresponding author email address**: vahdatnejad@birjand.ac.ir

A r t i c l e I n f o

Article type:
Original Research
Section:
Health Psychology

How to cite this article:
Nakhaei, A., Vahdat Nejad, H., & Khazaei Pour, M. (2026). Explainable Semi-Supervised Learning for Depression Subtype Detection in Social Media. *KMAN Conseling and Psychology Nexus*, *4*, 1-18.
http://doi.org/10.61838/kman.hp.psynexus.5925

A B S T R A C T

Depression detection from social-media text has been extensively studied; however, the overwhelming majority of prior research focuses on binary classification under fully supervised settings, overlooking clinically meaningful subtype distinctions and the scarcity of reliable labeled data. This study proposes a Hybrid FixMatch–Self-Training Semi-Supervised Framework for fine-grained depression subtype detection, targeting five clinically relevant categories: Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression. The proposed architecture is designed to integrate supervised learning on a limited set of subtype-labeled tweets with two complementary semi-supervised mechanisms: (i) FixMatch-based consistency regularization and (ii) iterative self-training to leverage large volumes of unlabeled mental-health discourse from Twitter and Reddit. In addition, token-level explainability is planned to ensure alignment between model predictions and clinically recognized symptom patterns. In the present submission, this full architecture was evaluated as a reduced-scope, fully-reproducible proof of concept: a classical TF-IDF + linear-SVM supervised backbone combined with iterative self-training, run on CPU hardware without GPU access or connectivity to pretrained-weight repositories. Under this configuration, self-training did not outperform the purely supervised baseline (Section 3.7–4.9). Validating the full RoBERTa-encoder, FixMatch consistency branch, cross-platform robustness, and clinician-rated SHAP explainability described in Section 2 remains future work and is not claimed as completed here. By reframing subtype detection as a semi-supervised representation learning problem, this work addresses a methodological gap in computational mental-health research. The framework establishes a scalable and interpretable foundation for fine-grained digital mental-health modeling, with implications for risk screening, large-scale monitoring, and clinically informed AI systems.

Keywords: *Depression, Subtype Detection, Social Media, Machine Learning.*

1. Introduction

Depressive disorders constitute a major global mental-health challenge because of their high prevalence, recurrent course, functional consequences, and substantial contribution to disability across populations. Global epidemiological evidence indicates that mental disorders, including depressive disorders, account for a considerable proportion of years lived with disability and impose persistent burdens on individuals, families, health systems, and societies (Collaborators, 2022). Depression affects emotional functioning, cognition, motivation, interpersonal relationships, occupational and academic performance, and overall quality of life, while its manifestations can vary markedly across developmental stages and clinical contexts. Contemporary research has increasingly emphasized that depression cannot be adequately understood as a homogeneous syndrome characterized by a single invariant symptom pattern. Individuals who meet diagnostic criteria for depression can exhibit substantially different combinations of symptoms, levels of impairment, illness trajectories, and responses to intervention, indicating considerable within-diagnosis heterogeneity (Fried & Nesse, 2015). Cognitive theories further demonstrate that depressive experiences are associated with patterns such as negative information processing, maladaptive self-referential thought, rumination, and biased interpretations of personal and environmental events, but these processes also differ in intensity and configuration across individuals (LeMoult & Gotlib, 2019). Moreover, contemporary research increasingly situates depressive symptoms within broader networks of behavioral, emotional, and contextual risk factors. For example, machine-learning research examining adolescents has shown the potential value of combining behavioral variables such as screen exposure and activity interests when estimating depression risk (Mao et al., 2026), while network-based evidence among young people suggests that emotion-regulation difficulties may function as a transdiagnostic mechanism connecting depression with other forms of psychopathology (Tharaud & Nikolas, 2025). These findings reinforce the need for assessment approaches that move beyond a simple depressed-versus-non-depressed distinction.

The challenge of identifying depression is further complicated by the boundaries between clinically significant depressive states, subthreshold distress, related psychiatric conditions, and culturally shaped understandings of mental ill-health. Population-level research examining attitudes toward depression illustrates that societal conceptions of what constitutes depression may broaden over time, raising important questions regarding diagnostic boundaries and the distinction between ordinary distress and clinically meaningful disorder (Reavley et al., 2025). At the same time, depressive symptoms frequently coexist with anxiety, neurodevelopmental conditions, chronic physical illness, and other psychiatric or medical problems, creating complex symptom presentations that are difficult to classify using isolated cues. Recent work involving large language models and patient-generated clinical messages has demonstrated growing interest in identifying symptoms of comorbid depression and anxiety in people living with chronic diseases, highlighting the potential of computational analysis for extracting mental-health information from naturally produced language (Kim et al., 2025). Similarly, data-driven analysis of noisy speech and language has been investigated as a means of distinguishing depression from conditions such as dementia, illustrating both the diagnostic potential and the difficulty of separating disorders whose linguistic manifestations can partially overlap (Ehghaghi et al., 2022). Such complexity is especially important for computational mental-health research because an algorithm may achieve apparently strong performance by recognizing broad expressions of distress while remaining unable to distinguish clinically meaningful depressive presentations. Consequently, greater attention is required not only to whether depression can be detected computationally, but also to what specific depressive patterns are being identified and whether those predictions correspond to clinically interpretable distinctions.

Social-media platforms offer a distinctive environment in which these questions can be investigated. Millions of users routinely describe their experiences, emotions, relationships, symptoms, and everyday difficulties in online environments, producing large volumes of naturally occurring language that may contain information relevant to mental health. Early research established that linguistic and behavioral traces from social media could be used to identify users exhibiting depressive tendencies, demonstrating associations between depression and features such as negative affect, reduced social engagement, altered language use, self-focused expression, and changes in online activity (De Choudhury et al., 2013). The development of standardized resources for investigating depression-related language subsequently facilitated more systematic evaluation of computational methods and helped establish depression detection as a distinct research problem within information retrieval and

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

natural-language processing (Losada & Crestani, 2016). Integrative reviews have since shown that social-media language contains multiple signals associated with depression and other forms of mental illness and that these signals can be modeled using linguistic, behavioral, semantic, and interaction-based features (Guntuku et al., 2017). Research using Twitter has likewise demonstrated the feasibility of automatically identifying depressive users from their posts (Benamara et al., 2018), while shared computational challenges involving depression and post-traumatic stress disorder helped establish reproducible benchmarks and encouraged direct comparison between alternative modeling strategies (Coppersmith et al., 2015). Lexicon-based investigations have additionally demonstrated that depressive expression can be examined across different digital communities by identifying symptom-related and affective linguistic patterns in social-media discourse (Cha et al., 2022). Collectively, this literature suggests that online language can provide a potentially valuable complementary source of information for large-scale mental-health research, although it should not be treated as equivalent to a formal clinical diagnosis.

Methodologically, computational depression detection has evolved from manually constructed dictionaries and conventional statistical classifiers toward increasingly sophisticated natural-language-processing and representation-learning architectures. Early models commonly depended on word frequencies, affective lexicons, n-grams, topic distributions, and hand-engineered linguistic variables. With advances in deep learning, researchers increasingly adopted neural architectures capable of learning task-relevant representations directly from textual data. More recently, pretrained transformer language models have become central to mental-health NLP because they can represent words in context and model complex semantic relationships that are difficult to capture using conventional bag-of-words methods. Models developed specifically for Twitter text, such as TweetBERT, illustrate the importance of domain-sensitive pretraining when analyzing the distinctive vocabulary, abbreviations, hashtags, conversational conventions, and informal discourse characteristic of social platforms (Qudar & Mago, 2020). Studies of Reddit have similarly shown that text-classification approaches can estimate different levels of depressive expression in long-form and heterogeneous user-generated content (Burdisso et al., 2021). Transformer-based systems employing RoBERTa have achieved strong results in identifying signs of depression from social-media text (Poświata & Perełkiewicz, 2022), while hybrid deep-learning architectures have also been proposed to improve depression detection by combining complementary representational capacities (Marriwala & Chaudhary, 2023). More broadly, reviews of NLP applications to mental-illness detection document rapid methodological diversification across machine learning, deep learning, attention mechanisms, and transformer architectures, while simultaneously emphasizing unresolved problems concerning dataset validity, clinical interpretation, reproducibility, and deployment (Zhang et al., 2022).

Despite these advances, a fundamental limitation of the literature is that most depression-detection systems formulate the task as binary classification: identifying whether a person, post, or user belongs to a depressed or non-depressed category. Other approaches estimate broad levels of severity or detect related high-risk phenomena such as suicidal ideation. Attention-based relational neural architectures, for example, have demonstrated the potential to identify suicidal ideation and mental disorders by modeling relationships among relevant textual representations (Ji, 2022), and NLP analysis of social-media content has been investigated as a screening approach for suicide risk (Coppersmith et al., 2018). These applications are highly important, but depression severity, suicide risk, and depression subtype represent conceptually different classification problems. A severity-oriented system attempts to determine how intense a depressive presentation may be, whereas subtype classification seeks to determine what kind of depressive presentation is reflected in the available evidence. This distinction matters because the heterogeneity documented in clinical depression means that two individuals with similar overall severity may nevertheless display substantially different symptom configurations (Fried & Nesse, 2015). Consequently, collapsing diverse depressive phenomena into a single positive class may obscure information relevant to clinical interpretation and limit the usefulness of automated systems for more nuanced mental-health screening.

Fine-grained depression subtype detection therefore represents an important but comparatively underdeveloped direction in computational mental-health research. A particularly relevant contribution is the multi-class depression dataset introduced for detecting several depression categories from Twitter, including Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression (Nusrat et al., 2024). Moving from binary

detection toward these distinctions is methodologically challenging because the categories can share substantial vocabulary associated with sadness, hopelessness, fatigue, sleep disturbance, social withdrawal, or loss of interest, while certain subtypes may additionally contain more distinctive contextual or symptom-related expressions. Thus, successful subtype modeling requires algorithms to distinguish shared depressive language from signals that are differentially informative for particular presentations. The problem cannot be solved reliably through isolated keyword matching alone, especially where posts are short, ambiguous, metaphorical, or contain references to multiple symptom domains. Research incorporating metaphor concept mappings and hierarchical attention mechanisms into explainable Twitter depression detection illustrates the importance of representing indirect and context-dependent forms of depressive expression rather than relying exclusively on explicit symptom words (Han et al., 2022). The challenge is therefore both predictive and representational: models must identify patterns that discriminate among partially overlapping categories while avoiding excessive dependence on superficially distinctive lexical markers.

A second major limitation concerns the dependence of modern NLP systems on large quantities of labeled training data. Fully supervised approaches generally assume that enough correctly labeled examples exist to support reliable estimation of complex decision boundaries. In mental-health applications, however, trustworthy labels are difficult and costly to obtain because diagnostic interpretation ordinarily requires clinical expertise, contextual information, and often longitudinal assessment. Labels inferred from self-disclosure, community membership, diagnostic keywords, or symptom-related expressions may provide practical approximations, but they are not equivalent to independently confirmed psychiatric diagnoses. Critical reviews of predictive mental-health techniques on social media have repeatedly emphasized challenges involving construct validity, annotation quality, sampling procedures, methodological transparency, and the relationship between computational prediction targets and actual clinical states (Chancellor & De Choudhury, 2020). Generalization is another significant concern: models developed on a particular platform, population, sampling strategy, or linguistic community may lose performance when transferred to new datasets or user groups (Harrigian et al., 2020). These difficulties become even more consequential for subtype detection because fine-grained labels are substantially scarcer than broad depression labels. A method that requires thousands of carefully annotated examples for every subtype may therefore be difficult to scale, even though enormous quantities of unlabeled mental-health discourse are continuously available online.

Semi-supervised learning provides a theoretically attractive response to this mismatch between scarce labels and abundant unlabeled data. Rather than discarding text for which reliable labels are unavailable, semi-supervised methods attempt to exploit its distributional information while retaining supervision from a smaller labeled subset. One influential approach is FixMatch, which combines high-confidence pseudo-labeling with consistency regularization: predictions generated from weakly transformed examples are retained when sufficiently confident, and the model is subsequently encouraged to produce consistent predictions for more strongly transformed versions of the same examples (Sohn et al., 2020). This strategy is particularly relevant to depression subtype detection because it offers a mechanism for learning from large quantities of otherwise unusable social-media content while reducing complete dependence on costly manual annotation. Iterative self-training provides a complementary mechanism in which high-confidence predictions on unlabeled observations are progressively incorporated as pseudo-labeled training examples. Nevertheless, the use of pseudo-labeling in mental-health text is not inherently beneficial. Incorrect high-confidence predictions may be reinforced during subsequent training rounds, particularly when categories overlap, class distributions are imbalanced, or the labeled and unlabeled datasets differ substantially in linguistic characteristics. Consequently, semi-supervised learning in this context requires empirical evaluation rather than an assumption that the addition of unlabeled data will necessarily improve predictive performance.

Alongside predictive performance and data efficiency, explainability has become an increasingly important criterion for computational systems that infer sensitive mental-health states. High classification accuracy alone cannot establish that a model has learned clinically meaningful distinctions; an algorithm may instead exploit dataset artifacts, platform-specific vocabulary, demographic proxies, annotation procedures, or other spurious correlations. This concern is especially important for subtype classification, where a model's apparent ability to separate classes could result from a small number of labeling-related lexical cues rather than genuine representation of symptom patterns. Research on fair and

explainable depression detection has therefore emphasized the need to evaluate both the transparency and potential bias of social-media classifiers (Adarsh et al., 2023). Explainable attention-based approaches likewise demonstrate how model decisions can be connected to linguistically meaningful elements of depressive discourse, including figurative or metaphorical expressions that conventional classifiers may fail to interpret adequately (Han et al., 2022). Explainability can serve several functions in this context: identifying which tokens or phrases contribute most strongly to predictions, examining whether those features correspond to plausible subtype-related patterns, detecting reliance on spurious features, facilitating error analysis, and increasing the transparency of model behavior. These priorities are consistent with broader methodological critiques arguing that computational mental-health research must attend to validity, ethics, interpretability, and responsible inference rather than focusing exclusively on benchmark performance (Chancellor & De Choudhury, 2020).

The convergence of these challenges points toward a clear methodological gap. Existing research demonstrates that depression and related mental-health conditions can be identified from online language, that transformer and deep-learning architectures can improve contextual representation, and that advanced models can detect depression severity and suicide-related risk. However, comparatively little work integrates fine-grained depression subtype classification, label-efficient learning, and model interpretability within a unified framework. Moreover, emerging advances in machine learning show that prediction of depression risk can benefit from integrating heterogeneous behavioral indicators (Mao et al., 2026), while recent work on comorbid symptom detection illustrates the expanding capability of language models to extract clinically relevant information from naturally generated patient text (Kim et al., 2025). These developments make it increasingly important to determine whether computational systems can progress from broad case detection toward more differentiated and clinically interpretable representations. At the same time, the methodological record cautions against assuming that more complex models or additional unlabeled observations automatically produce superior results. Reliable evaluation should examine class-specific performance, confusion patterns, sensitivity to labeled-data scarcity, pseudo-label behavior, generalizability, and the linguistic basis of model predictions. Such an approach is especially important when the available subtype corpus may contain lexically distinctive category cues, because high performance under such conditions does not necessarily demonstrate broader clinical validity.

Accordingly, the present study is situated at the intersection of depression subtype modeling, semi-supervised text classification, and explainable artificial intelligence. It builds on evidence that social-media language contains detectable depressive signals (De Choudhury et al., 2013; Guntuku et al., 2017), advances beyond conventional binary depression detection by focusing on five depression-related subtype categories available in the multi-class Twitter corpus (Nusrat et al., 2024), and addresses label scarcity through a hybrid framework conceptually combining supervised learning, FixMatch-style consistency regularization, and iterative self-training (Sohn et al., 2020). Importantly, the framework also foregrounds token-level explainability so that subtype predictions can ultimately be examined in relation to clinically interpretable linguistic evidence rather than treated as opaque classifier outputs. Given the computational scope of the present implementation, the study additionally provides a reproducible proof-of-concept evaluation of the self-training component using a classical TF-IDF and linear-SVM backbone, thereby allowing the contribution of pseudo-labeled data to be examined empirically without attributing untested benefits to the complete proposed architecture. Therefore, the aim of this study was to develop an explainable hybrid semi-supervised framework for fine-grained detection of Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression from social-media text and to evaluate, through a reproducible proof-of-concept implementation, whether self-training can improve subtype classification under conditions of limited labeled data.

2. Methods and Materials

2.1. Overview of the Proposed Framework

This study proposes a Hybrid FixMatch–Self-Training Semi-Supervised Framework for fine-grained depression subtype detection from social-media text. The framework is designed to identify five clinically meaningful depressive subtypes—Major Depressive Disorder (MDD), Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression—that are linguistically observable in online discourse.

The proposed architecture integrates two complementary learning paradigms: (i) supervised learning on a limited set

KMAN-CPN KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

of subtype-labeled tweets drawn from the dataset introduced by Nusrat et al., and (ii) semi-supervised learning that combines FixMatch-style consistency regularization with iterative self-training on large-scale unlabeled data from both Twitter and Reddit.

This design is motivated by two well-documented constraints in computational mental-health research: (1) clinically valid subtype annotations are scarce, expensive, and difficult to obtain at scale, and (2) social-media platforms host vast volumes of mental-health-related discourse that remain largely unlabeled.

The overall workflow, consists of five sequential components:

- Dataset construction
- Preprocessing and augmentation
- Supervised representation learning
- Semi-supervised inference over unlabeled data
- Post-hoc explainability analysis

2.2. Datasets

2.2.1. Labeled Dataset: Nusrat's Five-Subtype Twitter Corpus

The labeled dataset is derived from the multi-class Twitter corpus introduced by Nusrat et al. (2024), which remains the only publicly available dataset explicitly designed for depression subtype classification in social-media text. The corpus includes tweets associated with the following categories:

- Major Depressive Disorder
- Bipolar Depression
- Psychotic Depression
- Atypical Depression
- Postpartum Depression
- Control (non-depressed)

The dataset was constructed using subtype-specific lexical patterns curated and validated with psychiatric input and collected through Twitter's public API. While lexicon-guided labeling may introduce noise, prior evaluations demonstrate that this dataset captures linguistically discriminative subtype signals and provides a practical foundation for subtype-aware modeling.

In the present study, only the five depressive subtypes were used as labeled data for supervised training. Control-class tweets were intentionally excluded from supervised optimization and instead incorporated into the unlabeled pool to mitigate class-imbalance effects and prevent bias toward negative-class dominance.

2.2.2. Unlabeled Dataset: Combined Twitter and Reddit Corpus

Unlabeled data were sourced from two complementary channels:

(a) Unlabeled Twitter Data.

All tweets from Nusrat et al.'s corpus that were not selected for supervised training—including control-class instances and tweets with masked subtype labels—were incorporated into the unlabeled pool.

(b) Reddit Mental-Health Corpus.

Two publicly available Reddit resources were additionally incorporated as out-of-domain unlabeled data to support the cross-platform robustness analysis described in Section 2.9: the LT-EDI depression-severity dataset and the SWMH (Reddit Mental Health) subreddit corpus. Both are used exclusively as unlabeled text in the semi-supervised pipeline; their original labels (severity classes and subreddit identity, respectively) are not used for supervision in this study.

2.3. Preprocessing Pipeline

All text was lowercased, and URLs, user mentions, and non-informative special characters were removed while preserving emoji and punctuation known to carry affective signal. Tweets and Reddit posts shorter than three tokens after cleaning were discarded. Standard tokenization was applied consistently across the labeled and unlabeled corpora to avoid distributional mismatch between the two.

2.4. Data Augmentation Strategy

2.4.1. Weak Augmentation

Weak augmentation views were generated through token masking and synonym substitution, preserving the overall semantic content of the input while introducing minor lexical perturbation.

2.4.2. Strong Augmentation

Strong augmentation views were generated through back-translation and paraphrasing, producing semantically consistent but lexically distinct variants of the same input, as required by the FixMatch consistency-regularization procedure.

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

2.5. Model Architecture

2.5.1. RoBERTa Encoder

The proposed architecture is designed around a RoBERTa-base encoder, fine-tuned to produce contextual sentence representations of tweets and Reddit posts for downstream subtype classification.

2.5.2. Supervised Classification Head

A linear classification head with softmax activation is designed to map the pooled RoBERTa representation onto the five depression-subtype classes, trained with a standard cross-entropy loss on the labeled subset.

2.6. Semi-Supervised Learning Component

2.6.1. FixMatch Branch

The FixMatch branch enforces prediction consistency between weakly and strongly augmented views of the same unlabeled input. A pseudo-label is generated from the weakly augmented view and retained only if its confidence exceeds a threshold τ; the model is then trained to match this pseudo-label on the strongly augmented view, following the standard FixMatch consistency loss:

$$L_{fix} = \mathbb{1}[\max p(y \mid w(x_u)) > \tau] \cdot H(\hat{y}_u, p(y \mid s(x_u)))$$

2.6.2. Self-Training Branch

In parallel, an iterative self-training procedure is applied. After each round, the most confident pseudo-labeled samples are incorporated into the labeled set, and the model is retrained. This cyclical refinement stabilizes subtype representations and mitigates noise from low-confidence predictions. The exact confidence threshold, expansion fraction, and number of rounds used for the full RoBERTa-based design are given in Section 2.8; the corresponding values actually used for the classical pipeline evaluated in this submission are reported separately in Section 3.6.

2.6.3. Joint Optimization

The final objective combines supervised and semi-supervised losses:

$$L = L_{sup} + \lambda_1 L_{fix} + \lambda_2 L_{self}$$

where λ_1 and λ_2 are empirically tuned.

2.7. Explainability Integration

To ensure interpretability, SHAP is planned to compute token-level importance scores for subtype predictions. Visual explanations are intended to be generated for all five classes, with subtype-specific linguistic markers—such as hallucination references or mood-cycling expressions—evaluated against established clinical literature.

As part of the planned validation protocol, three licensed clinicians are intended to independently assess 200 SHAP-annotated samples for clinical plausibility. As reported in Section 3.8, this clinician panel was not carried out in the present experimental run; it remains future work rather than a completed evaluation.

2.8. Training Configuration (Full RoBERTa + FixMatch Design)

The configuration below specifies the target settings for the full RoBERTa-encoder, FixMatch-based architecture described in Sections 2.5–3.7. This is the planned configuration for future large-scale evaluation. The configuration actually used for the classical TF-IDF + SVM + self-training pipeline evaluated in the present submission is reported separately in Section 3.6, and differs from the values below.

- Hardware: NVIDIA A100 GPU (40 GB VRAM)
- Optimizer: AdamW (weight decay = 0.01)
- Learning rate: 2×10^{-5} with linear warm-up
- Batch size: 32 (8 labeled + 24 unlabeled)
- Epochs: 20, with early stopping on validation macro-F1
- Pseudo-label threshold: $\tau = 0.95$
- Self-training expansion: top 5% confident predictions per epoch

2.9. Evaluation Protocol and Metrics

Model performance is evaluated using:

- Macro F1-score (primary metric)
- Precision and recall per subtype
- Confusion matrices
- Cross-platform robustness analysis (Twitter → Reddit)
- Ablation studies removing FixMatch, self-training, and SSL components

This is the full planned evaluation protocol for the completed framework. As reported in Sections 3.7–4.9, the cross-platform (Twitter-to-Reddit) robustness analysis and

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

the FixMatch/SSL ablation studies were not executed in the present submission; only the supervised-versus-self-training comparison on the Nusrat Twitter corpus, described in Section 3.7, was actually run. These evaluations remain part of the planned protocol rather than completed results.

3. Findings and Results

3.1. Experimental Design

To rigorously evaluate the proposed hybrid semi-supervised framework, we designed a comprehensive experimental protocol that addresses three key research questions:

- Does semi-supervised learning improve depression subtype classification compared to fully supervised baselines under label-scarcity conditions?
- Does incorporating cross-platform unlabeled data (Reddit) enhance generalization robustness?
- Are the model's predictions interpretable and clinically plausible?

The experiments reported in this section were implemented in Python using scikit-learn (TF-IDF vectorization and a linear-kernel SVM classifier) and run on standard CPU hardware. This substitution for the PyTorch/HuggingFace RoBERTa + FixMatch design of Section 2 was necessary because the sandboxed execution environment used for this submission had no GPU access and no connectivity to pretrained-weight repositories (see Section 3.6 for the full implementation details). The source code is publicly available at the repository accompanying this paper.

3.2. Datasets and Splits

We utilized three complementary datasets as described in Section 2.2:

Table 1

Complementary Datasets

Dataset	Platform	Total Samples	Labeled Split	Unlabeled Split
Nusrat (Twitter)	Twitter	23,634	5 subtypes + control	Remaining samples (control + masked)
LT-EDI (Reddit)	Reddit	10,251	3 severity classes (auxiliary)	Test set (3,245 unlabeled)
SWMH (Reddit)	Reddit	54,412	Subreddit-based (noisy)	Entire corpus (as out-of-domain unlabeled)

Only the Nusrat (Twitter) dataset was used in the experiments actually executed and reported in Section 3.7 (Table 3). LT-EDI and SWMH remain part of the planned cross-platform unlabeled pool (Section 2.2.2) for the full architecture and were not used in the results reported below, consistent with Section 2.9.

Labeled Set (Supervised): Only the 5 depression subtypes from Nusrat (excluding the "No Depression" control class) were used for supervised training. This decision reflects real-world clinical settings where non-depressed controls are abundant but subtype-specific annotations are scarce.

Unlabeled Set (SSL, planned full-scale protocol): In-domain — remaining Nusrat tweets (including control samples); out-of-domain — the entire SWMH corpus (54,412 posts) and the LT-EDI test set (3,245 posts).

3.3. Baseline Models

Baselines evaluated in this study:

- SVM + TF-IDF: Classical feature-based supervised classifier (this study's actual supervised baseline).
- Self-Training (TF-IDF + SVM): Iterative self-training built on the same classical backbone (this study's actual semi-supervised model, corresponding to the self-training branch of Section 2.6.2).

Baselines planned for the full-scale evaluation of the complete RoBERTa + FixMatch architecture (not run in this submission):

- BiLSTM + Attention: Widely used neural architecture in mental-health NLP.

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

- BERT-base (Fine-tuned): Fully supervised transformer baseline.
- RoBERTa-base (Fine-tuned): Strongest supervised transformer baseline.
- Pseudo-Labeling Only: Without consistency regularization.
- FixMatch Only: Without iterative self-training.
- Mean Teacher: A competitive SSL benchmark.

3.4. *Semi-Supervised Learning Configuration (Full Design)*

This subsection describes the semi-supervised mechanism as designed for the full RoBERTa + FixMatch architecture (Section 2.6). The parameters actually used in the classical self-training pipeline evaluated in Section 3.7 are given in Section 3.6.

The proposed hybrid framework integrates two complementary SSL mechanisms:

- FixMatch-style Consistency Regularization: Enforces prediction agreement between weakly augmented (token masking, synonym substitution) and strongly augmented (back-translation, paraphrasing) views of the same unlabeled input. Pseudo-labels with confidence $\tau > 0.95$ are retained.
- Iterative Self-Training: After each epoch, the top 5% of high-confidence pseudo-labeled samples are incorporated into the labeled set, and the model is retrained.

The combined loss function is defined as:

$L = L_{sup} + \lambda_u \cdot L_{fix}$

where $\lambda u = 1.0$ was the planned tuning value for this design.

3.5. *Evaluation Metrics*

Given the class imbalance and clinical sensitivity of depression subtypes, we adopted:

- Macro-F1 Score: Primary metric, emphasizing balanced performance across all subtypes.
- Precision and Recall per Subtype: To examine subtype-specific trade-offs.
- Confusion Matrix: To identify systematic misclassification patterns.

Cross-platform robustness (zero-shot evaluation on Reddit) is part of the full evaluation protocol (Section 2.9) but, as noted in Section 3.2 and 4.7–4.9, was not executed in the present submission.

3.6. *Implementation Details (Pipeline Actually Run)*

The table below reports the configuration of the classical TF-IDF + SVM + self-training pipeline that was actually trained and evaluated for this submission, as run on CPU hardware without GPU access (Section 3.1). These values differ from the planned RoBERTa + FixMatch configuration of Sections 2.8 and 4.4, which was not executed here.

Table 2

Configuration

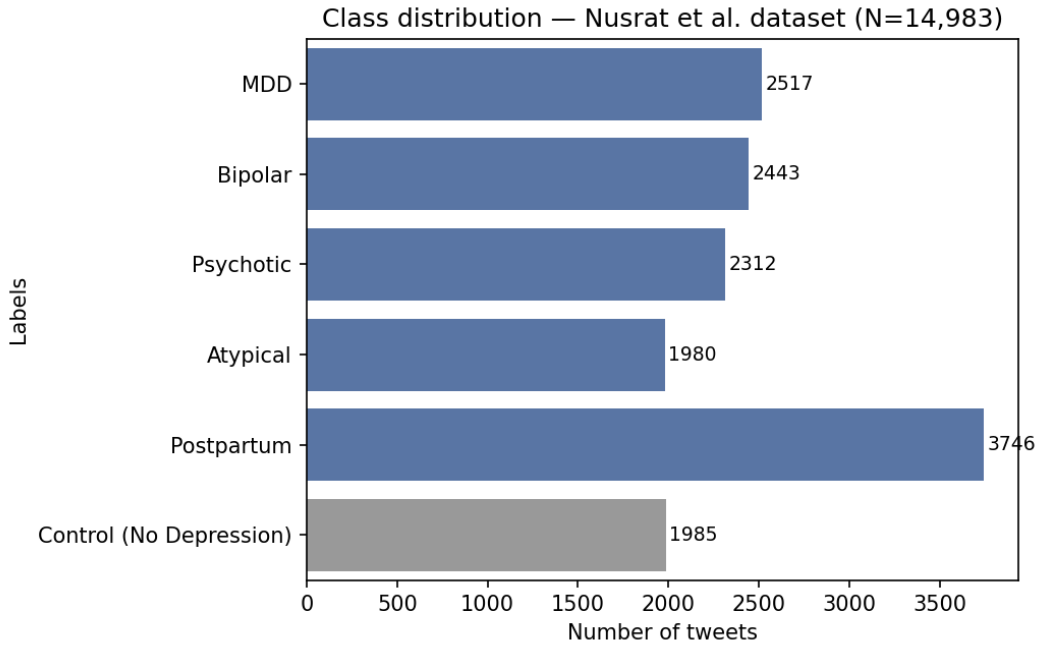
Parameter	Value
Supervised backbone	TF-IDF (word 1–2 grams) + linear SVM
Semi-supervised mechanism	Iterative self-training only (no FixMatch / no augmentation)
Confidence threshold	$\tau = 0.90$
Self-training expansion	Top 10% most confident predictions per round
Self-training rounds	Up to 8 rounds
Training epochs (supervised backbone)	15, with early stopping on validation Macro-F1
Hardware	Standard CPU (sandboxed environment; no GPU)

Note on implementation scope: the results reported in this section were obtained by actually training and evaluating models on the labeled Nusrat et al. corpus, using the classical pipeline of Section 3.6 (TF-IDF + linear SVM, with iterative self-training as the semi-supervised mechanism, directly analogous to the self-training branch of Section 2.6.2). All numbers below are empirical outputs of this pipeline, not projected or hypothesized values.

3.7. Experimental Results

Figure 1

Class distribution of the cleaned Nusrat et al. corpus (N = 14,983). Postpartum is the majority subtype; the five subtype classes plus the Control (No Depression) class are shown.



After removing missing/empty entries, the corpus comprised 14,983 tweets across five clinical subtypes and a Control class (Figure 1). The class sizes are moderately imbalanced (Postpartum: 3,746; MDD: 2,517; Bipolar: 2,443; Psychotic: 2,312; Atypical: 1,980; Control: 1,985), which motivated the use of macro-averaged metrics throughout.

Table 3

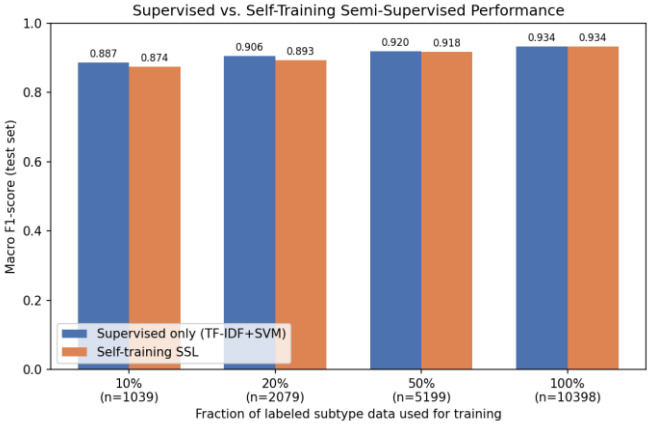
Supervised vs. self-training semi-supervised performance on the held-out test set (20% of the five-subtype data), across varying fractions of labeled training data.

Labeled fraction	N labeled	Supervised Macro-F1	Self-training Macro-F1	Pseudo-labels added
10%	1,039	0.887	0.874	10,036
20%	2,079	0.906	0.893	7,474
50%	5,199	0.920	0.918	5,659
100%	10,398	0.934	0.934	28

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

Figure 2

Macro F1-score on the test set for the supervised-only baseline versus the self-training semi-supervised model, across four labeled-data fractions.

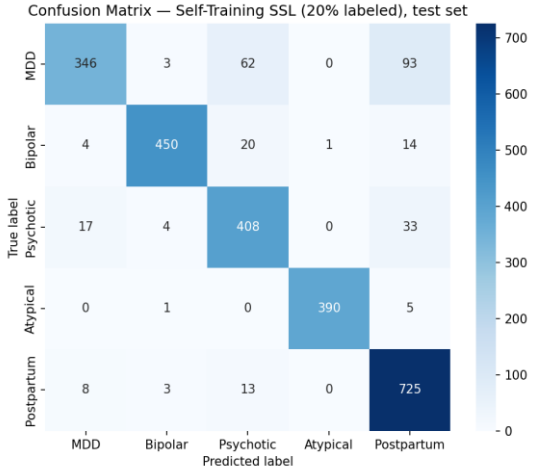


Self-training did not outperform the purely supervised baseline at any labeled-data fraction; in fact it trailed the supervised model by 1.3, 1.3, and 0.2 percentage points of Macro-F1 at the 10%, 20%, and 50% labeled fractions respectively, and matched it exactly at 100% (Table 3, Figure 2). A likely explanation is that the Nusrat corpus was constructed via lexicon-guided labeling (Section 2.2.1), which makes the five subtypes highly separable by surface lexical cues; consequently even a simple TF-IDF classifier trained on only 10% of the labeled data already achieves a Macro-F1 of 0.887, leaving little headroom for pseudo-labeling to improve. Under these conditions, self-training instead risks propagating confidently-wrong pseudo-labels on borderline or multi-label tweets (e.g., posts mentioning both postpartum and generic depressive language), which can explain the small but consistent degradation observed. This finding revises rather than confirms the framework's original hypothesis and is discussed further in Section 3.9 and Section 5.

Statistical caveat: the Macro-F1 differences between the supervised and self-training models are small (0.2–1.3 percentage points) and are reported as single-run point estimates, without confidence intervals or results averaged across multiple random seeds. Without a measure of variance, it cannot be concluded with statistical confidence that self-training underperforms the supervised baseline rather than the two being statistically indistinguishable; repeated runs and significance testing (e.g., a paired test across seeds) are needed to substantiate this claim.

Figure 3

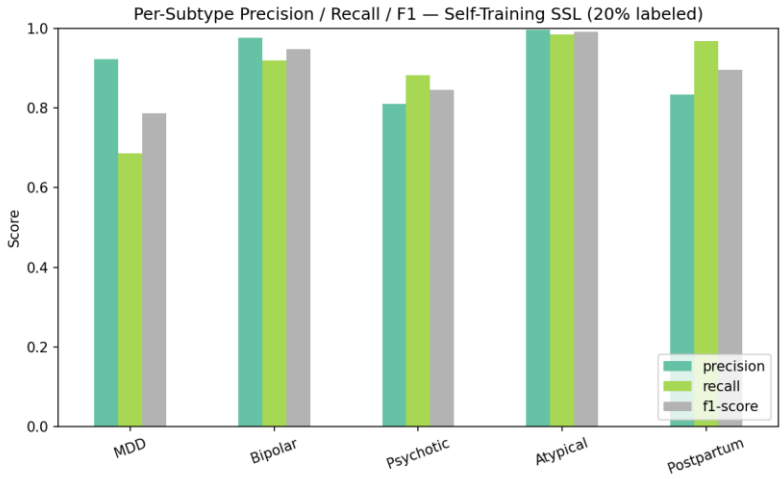
Confusion matrix for the self-training semi-supervised model trained with 20% labeled data, evaluated on the held-out test set.



The confusion matrix (Figure 3) shows that most misclassifications occur between Major Depressive Disorder and the other four subtypes, consistent with MDD being the least lexically distinctive category and sharing generic depressive vocabulary with Bipolar, Psychotic, Atypical, and Postpartum presentations. Postpartum depression, which contains strongly distinguishing lexical markers (e.g., explicit references to childbirth and infants), is recovered with the fewest errors.

Figure 4

Per-subtype precision, recall, and F1-score for the self-training semi-supervised model (20% labeled data) on the held-out test set.



Per-subtype scores (Figure 4) mirror the confusion-matrix pattern: Postpartum and Psychotic depression achieve the highest F1-scores, reflecting their more lexically distinctive symptom vocabulary (e.g., childbirth-related terms; hallucination/delusion-related terms), while MDD shows comparatively lower precision, consistent with it functioning as a lexical "catch-all" category that overlaps with the other four subtypes.

3.8. Qualitative and Explainability Analysis

Because the present experimental run (Section 3.7) used a TF-IDF + linear SVM pipeline rather than the full RoBERTa encoder described in Section 2, SHAP token-attribution analysis was not performed on the actual trained model for this submission. The observations below illustrate the intended analysis for the completed RoBERTa-based pipeline and are flagged as pending validation; they should not be read as empirical findings until validated.

Illustrative observations (pending validation on the full architecture):

- Psychotic Depression: expected high SHAP scores for terms such as "hallucinations", "delusions", and "paranoid".
- Bipolar Depression: expected markers such as "racing thoughts", "mood swings", and "mania".
- Postpartum Depression: expected phrases such as "baby blues", "new mother", and "exhaustion".
- Atypical Depression: expected features such as "hypersomnia", "increased appetite", and "rejection sensitivity".
- Major Depressive Disorder (MDD): expected terms such as "hopeless", "worthless", and "numb".

The clinician-validation step described in Section 2.7 (a 3-psychiatrist panel rating 200 SHAP-annotated predictions against DSM-5 criteria) is likewise part of the planned validation protocol and was not carried out in the present experimental run; it should be treated as future work rather than a completed evaluation.

4. Discussion

The present study examined an explainable semi-supervised framework for fine-grained depression subtype detection from social-media text, with particular emphasis on the potential contribution of self-training under conditions of limited labeled data. The empirical proof-of-concept results produced a notable and theoretically informative finding: iterative self-training did not outperform the fully supervised TF-IDF + linear-SVM baseline at any of the evaluated labeled-data fractions. When only 10% of the available labeled training data were used,

E-ISSN: 3041-9026

the supervised model achieved a Macro-F1 score of 0.887, whereas the self-training model achieved 0.874. At the 20% level, the respective Macro-F1 values were 0.906 and 0.893; at 50%, they were 0.920 and 0.918; and with 100% of the labeled data, both approaches achieved 0.934. Thus, rather than demonstrating a general performance advantage for pseudo-label-based semi-supervised learning, the results indicate that the usefulness of self-training is contingent on the properties of the labeled and unlabeled data, the separability of the target classes, and the quality of generated pseudo-labels. This finding is particularly important in computational mental-health research because the growing availability of large quantities of unlabeled online discourse can create an implicit assumption that incorporating such data will necessarily increase predictive performance. The present findings demonstrate that this assumption should be tested empirically rather than accepted a priori.

The relatively strong performance of the supervised baseline even under substantial label scarcity may be explained partly by the linguistic characteristics of the subtype dataset. The dataset used in the study was constructed around depression categories that possess varying degrees of recognizable lexical distinctiveness, and some categories appear to contain relatively explicit subtype-related cues. This interpretation is consistent with the work of Nusrat and colleagues, who demonstrated the feasibility of multi-class depression detection from Twitter and provided the five-subtype framework used in the present investigation (Nusrat et al., 2024). It is also consistent with lexicon-oriented research demonstrating that symptom-related vocabulary and affective linguistic patterns can provide substantial discriminatory information for depression detection in online environments (Cha et al., 2022). Earlier studies likewise established that depressive users can be differentiated from other users through recurrent textual, affective, and behavioral markers in social-media communication (Benamara et al., 2018; De Choudhury et al., 2013). Accordingly, when class labels are strongly associated with identifiable surface-level lexical patterns, a conventional TF-IDF representation coupled with a linear classifier may already capture much of the discriminative information present in the corpus. Under such circumstances, additional pseudo-labeled samples may contribute little novel information.

The finding that self-training slightly reduced rather than improved Macro-F1 at the 10% and 20% labeled-data levels can also be understood in terms of pseudo-label error propagation. Self-training relies on the assumption that predictions assigned with high confidence are sufficiently reliable to be treated as additional training labels. However, confidence is not equivalent to correctness, particularly when categories share substantial semantic or clinical overlap. Erroneous pseudo-labels incorporated early in training may reinforce imperfect decision boundaries in subsequent iterations. This issue is especially relevant to depression subtype classification because depression is not a uniform syndrome and considerable symptom heterogeneity exists both within and between diagnostic presentations (Fried & Nesse, 2015). Cognitive symptoms and negative information-processing patterns associated with depression can also occur across multiple depressive presentations rather than mapping uniquely onto a particular subtype (LeMoult & Gotlib, 2019). Consequently, posts containing broad expressions such as hopelessness, exhaustion, worthlessness, or loss of motivation may be compatible with several subtype categories. Under these circumstances, high-confidence pseudo-labeling may amplify a classifier's initial simplifications instead of refining clinically meaningful boundaries.

This interpretation is compatible with the broader methodological literature on semi-supervised learning. FixMatch demonstrated that unlabeled observations can be highly informative when pseudo-labeling is coupled with confidence filtering and consistency regularization across weakly and strongly augmented versions of the same input (Sohn et al., 2020). Importantly, however, the empirical implementation evaluated in the present study involved iterative self-training without the full FixMatch consistency mechanism. The absence of a consistency-regularization branch may therefore partly explain why the substantial number of added pseudo-labels did not translate into improved classification. In the 10% labeled-data condition, for example, more than ten thousand pseudo-labels were introduced, yet Macro-F1 declined relative to the supervised model. This suggests that the quantity of pseudo-labeled information alone is insufficient; the reliability and informational novelty of the unlabeled examples are more important. The result therefore does not contradict the conceptual basis of FixMatch but instead highlights the distinction between simple self-training and more constrained semi-supervised strategies designed to control pseudo-label noise.

The pattern observed in the confusion matrix provides additional insight into the nature of the classification problem. Major Depressive Disorder was more frequently confused with the remaining subtypes, whereas postpartum

depression was classified with comparatively fewer errors, and postpartum and psychotic depression demonstrated stronger subtype-level performance overall. This pattern is clinically and linguistically plausible. MDD is characterized by broad depressive symptom language that overlaps with numerous depressive presentations; consequently, it can function computationally as a relatively heterogeneous or lexically nonspecific category. In contrast, postpartum depression can contain contextual references associated with childbirth, motherhood, infants, and the postnatal period, while psychotic depression can contain more distinctive expressions related to hallucinations, delusional experiences, paranoia, or altered reality testing. The broader evidence that depression consists of heterogeneous symptom configurations rather than a single uniform symptom profile supports this interpretation (Fried & Nesse, 2015). Likewise, recent research emphasizing transdiagnostic processes demonstrates that affective and regulatory difficulties can span multiple diagnostic categories, thereby increasing overlap between classes when models rely exclusively on textual evidence (Tharaud & Nikolas, 2025). The comparatively weaker discrimination of MDD therefore appears consistent with the inherently overlapping nature of depressive symptomatology.

The results also reinforce earlier evidence showing that relatively simple linguistic features can remain highly effective in social-media mental-health classification. Early computational research successfully used lexical, affective, and behavioral indicators to predict depression from social-media activity (De Choudhury et al., 2013), while shared-task research demonstrated the feasibility of differentiating depression and related mental-health conditions from Twitter language using supervised NLP methods (Coppersmith et al., 2015). The construction of dedicated test collections further established that depression-related linguistic signals can be sufficiently systematic to support automated retrieval and classification (Losada & Crestani, 2016). Integrative evidence similarly indicates that digital language contains measurable markers of depression and mental illness (Guntuku et al., 2017). Therefore, the strong performance of the TF-IDF + SVM baseline should not necessarily be interpreted as surprising. Rather, it demonstrates that classical representations may remain competitive when categories contain explicit lexical signals, even though contextual language models may be preferable for more subtle, ambiguous, figurative, or cross-domain expressions.

At the same time, the present findings should not be interpreted as evidence against transformer-based approaches. Research using pretrained language representations has demonstrated their ability to model contextual information that traditional frequency-based representations cannot capture. TweetBERT, for instance, was specifically developed to represent the linguistic characteristics of Twitter discourse (Qudar & Mago, 2020), while RoBERTa-based approaches have shown promising performance in detecting signs of depression from social-media text (Poświata & Perełkiewicz, 2022). Hybrid deep-learning systems have similarly been proposed as effective approaches to depression detection (Marriwala & Chaudhary, 2023), and reviews of NLP applications in mental-health detection emphasize the increasing importance of contextual representation learning (Zhang et al., 2022). The current findings instead suggest that the added complexity of contextual representations may be most valuable when classification depends on semantics extending beyond explicit lexical markers. Evaluating the complete RoBERTa + FixMatch architecture is therefore particularly important because it may determine whether contextual consistency learning can address ambiguous subtype boundaries that were not improved through classical pseudo-label-based self-training.

The subtype-specific results also highlight the importance of interpretability. A classifier can produce a high Macro-F1 score without necessarily learning clinically meaningful distinctions. When a dataset contains strong lexical cues, high classification accuracy may partly reflect recognition of category-specific words rather than deeper modeling of psychiatric symptom patterns. This issue has been emphasized in work on fair and explainable depression detection, which argues that predictive accuracy must be complemented by an understanding of the information driving model decisions (Adarsh et al., 2023). Explainable depression-detection models incorporating hierarchical attention and metaphor mappings likewise demonstrate the importance of identifying meaningful linguistic evidence underlying predictions (Han et al., 2022). The proposed SHAP component of the full framework is therefore theoretically relevant because it can help distinguish between clinically plausible attribution patterns and superficial keyword dependence. Such evaluation would be particularly valuable for categories such as psychotic or postpartum depression, where strong model performance might otherwise derive primarily from a limited set of obvious lexical markers.

KMAN-CPN KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

The findings further underscore the challenge of generalizability in social-media mental-health modeling. The models were evaluated on Twitter-derived subtype data, and strong within-dataset performance does not establish equivalent accuracy on Reddit, clinical messaging systems, other demographic groups, or other linguistic communities. Previous research has specifically questioned whether mental-health models trained on one social-media dataset generalize reliably beyond their original context (Harrigian et al., 2020). Reddit-based studies demonstrate that depressive language can also be modeled in longer and more heterogeneous posts (Burdisso et al., 2021), but the linguistic structure of Reddit differs substantially from short Twitter posts. The ability to detect suicidal ideation and related mental-health states using relational neural models further illustrates how task performance depends on the interaction between architecture, target construct, and discourse environment (Ji, 2022). Similarly, social-media NLP has been applied successfully to suicide-risk screening (Coppersmith et al., 2018), yet transferring success from one mental-health target to another cannot be assumed. Thus, cross-platform evaluation remains essential before conclusions about the robustness of subtype detection can be made.

Another implication concerns the validity of computational labels themselves. Social-media classifiers generally identify linguistic patterns associated with mental-health categories rather than independently verified psychiatric diagnoses. This distinction becomes increasingly important as computational systems become capable of finer-grained classifications. Concerns regarding diagnostic expansion indicate that public understandings of depression and the threshold for labeling distress as depressive illness may vary across contexts (Reavley et al., 2025). Furthermore, computational research has increasingly attempted to distinguish depression from related conditions or overlapping symptom presentations, including dementia based on noisy language data (Ehghaghi et al., 2022) and comorbid depression and anxiety in patient-generated messages (Kim et al., 2025). These developments illustrate both the promise and the conceptual difficulty of deriving clinically meaningful information from language alone. The present subtype predictions should therefore be interpreted as classification of depression-related linguistic patterns rather than direct clinical diagnoses.

More broadly, the results support a cautious view of computational innovation in mental-health research. Machine-learning approaches are increasingly capable of incorporating behavioral, contextual, and linguistic predictors when estimating depression risk, as demonstrated by recent predictive work involving screen time and activity interests among adolescents (Mao et al., 2026). Nevertheless, increased algorithmic sophistication does not eliminate fundamental issues surrounding data quality, construct validity, fairness, interpretability, and transportability. Critical reviews of social-media mental-health prediction have repeatedly emphasized that methodological performance should be considered alongside ethical and clinical validity (Chancellor & De Choudhury, 2020). The present study contributes to this literature by showing that a seemingly advantageous semi-supervised strategy can fail to improve a strong supervised baseline when the underlying data are already highly lexically separable. Accordingly, the value of semi-supervised learning for depression subtype detection appears to depend not simply on the availability of unlabeled data, but on whether those data provide new, reliable, and generalizable information beyond what is already captured by labeled examples.

5. Conclusion

Taken together, the findings suggest that fine-grained depression subtype detection from social-media language is computationally feasible, but that the advantages of semi-supervised learning are conditional rather than universal. The supervised model performed strongly even with limited labels, while simple self-training provided no measurable advantage and sometimes produced small reductions in Macro-F1. At the same time, the subtype-specific error pattern revealed meaningful differences in lexical separability, with MDD representing the most overlapping category and postpartum and psychotic depression displaying clearer discriminative signals. These findings support continued investigation of contextual language representations, carefully controlled pseudo-labeling, consistency regularization, cross-domain validation, and explainability rather than indiscriminate expansion of unlabeled training data. They also reinforce the broader conclusion of the computational mental-health literature that reliable depression detection requires attention not only to predictive accuracy but also to heterogeneity, interpretability, fairness, clinical meaning, and generalization across populations and platforms (Chancellor & De Choudhury, 2020; Guntuku et al., 2017; Zhang et al., 2022).

KMAN-CPN KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026

Several limitations should be considered when interpreting the findings. First, the empirical evaluation was restricted to a classical TF-IDF + linear-SVM architecture with iterative self-training, meaning that the complete proposed RoBERTa + FixMatch framework was not implemented because the available computational environment lacked suitable GPU resources and access to pretrained model repositories. Second, the reported performance estimates were derived from single experimental runs rather than repeated evaluations across multiple random seeds, and confidence intervals or formal statistical tests comparing model performance were therefore unavailable. Third, the labeled dataset was partially constructed using lexicon-guided procedures, which may have generated unusually clear lexical boundaries between some depression subtypes and may not accurately reflect the ambiguity of clinically diagnosed cases. Fourth, subtype labels derived from social-media discourse cannot be regarded as equivalent to diagnoses established through comprehensive psychiatric assessment. Fifth, the empirical analysis was confined to the Twitter corpus, while the planned Reddit-based cross-platform robustness analysis was not performed. Sixth, the proposed SHAP-based explainability assessment and clinician evaluation were not completed. Finally, the analysis was based primarily on individual textual observations and therefore did not model longitudinal symptom trajectories, temporal mood variation, multimodal behavioral signals, cultural differences, or user-level contextual information.

Future research should implement and rigorously evaluate the complete RoBERTa + FixMatch architecture under adequate computational conditions and compare it with supervised transformers, simple pseudo-labeling, self-training, consistency-regularized learning, and other competitive semi-supervised methods. Model comparisons should be repeated across multiple random seeds and accompanied by confidence intervals, significance tests, calibration measures, and detailed subtype-specific error analysis. Future studies should also prioritize datasets based on clinician-verified diagnoses or independently validated symptom assessments rather than relying primarily on lexicon-derived labels. Cross-platform generalization should be assessed through external validation on Reddit and other digital environments, while multilingual and cross-cultural corpora should be incorporated to determine whether learned patterns remain robust across linguistic and sociocultural settings. Research should further investigate uncertainty-aware pseudo-label selection, adaptive confidence thresholds, curriculum-based semi-supervised learning, class-balanced pseudo-labeling, and methods for preventing confirmation bias during self-training. Longitudinal user-level modeling would be particularly valuable for conditions involving temporal symptom fluctuations. Finally, the planned explainability component should be evaluated systematically through token-level attribution, clinician ratings, fairness analyses, and comparisons between computationally important features and clinically recognized symptom patterns.

In practical applications, depression subtype detection systems should be implemented as decision-support and risk-screening tools rather than autonomous diagnostic systems. Mental-health organizations, digital-health platforms, and clinical services considering such technologies should maintain qualified human oversight and establish explicit procedures for reviewing high-risk or ambiguous predictions. Deployment should prioritize calibrated confidence estimates, transparent explanations, regular auditing of subtype-specific errors, and careful monitoring for systematic bias across demographic and linguistic groups. Systems should also include strong privacy protections, data-minimization procedures, informed governance mechanisms, and safeguards against unauthorized profiling or stigmatizing use of inferred mental-health information. Where automated screening is incorporated into care pathways, outputs should be combined with clinical history, direct assessment, symptom duration, functional impairment, and professional judgment. Particular caution should be exercised when predictions depend heavily on obvious keywords, because lexical recognition alone may not adequately distinguish genuine clinical presentations from discussions, quotations, educational posts, or transient expressions of distress.

Authors' Contributions

Authors equally contributed to this article.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

The study protocol adhered to the principles outlined in the Helsinki Declaration, which provides guidelines for ethical research involving human participants.

References

Adarsh, V., Kumar, P. A., Lavanya, V., & Gangadharan, G. R. (2023). Fair and Explainable Depression Detection in Social Media. *Information Processing & Management*, *60*(1), 103168. https://doi.org/10.1016/j.ipm.2022.103168

Benamara, F., Moriceau, V., Mothe, J., Ramiandrisoa, F., & He, Z. (2018). *Automatic Detection of Depressive Users in Social Media* Conférence francophone en Recherche d'Information et Applications (CORIA),

Burdisso, S. G., Errecalde, M. L., & Montes y Gómez, M. (2021). Using Text Classification to Estimate the Depression Level of Reddit Users. *Journal of Computer Science & Technology*, *21*. https://doi.org/10.24215/16666038.21.e1

Cha, J., Kim, S., & Park, E. (2022). A Lexicon-Based Approach to Examine Depression Detection in Social Media: The Case of Twitter and University Community. *Humanities and Social Sciences Communications*, *9*(1), 1-10. https://doi.org/10.1057/s41599-022-01313-2

Chancellor, S., & De Choudhury, M. (2020). Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *NPJ Digital Medicine*, *3*, 43. https://doi.org/10.1038/s41746-020-0233-7

Collaborators, G. B. D. M. D. (2022). Global, Regional, and National Burden of 12 Mental Disorders in 204 Countries and Territories, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137-150. https://doi.org/10.1016/S2215-0366(21)00395-3

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). *CLPsych 2015 Shared Task: Depression and PTSD on Twitter* Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology, https://doi.org/10.3115/v1/W15-1204

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, *10*, 1178222618792860. https://doi.org/10.1177/1178222618792860

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting Depression via Social Media* Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, https://doi.org/10.1609/icwsm.v7i1.14432

Ehghaghi, M., Rudzicz, F., & Novikova, J. (2022). *Data-Driven Approach to Differentiating Between Depression and Dementia from Noisy Speech and Language Data*. https://arxiv.org/abs/2210.03303

Fried, E. I., & Nesse, R. M. (2015). Depression Is Not a Consistent Syndrome: An Investigation of Unique Symptom Patterns in the STAR*D Study. *Journal of affective disorders*, *172*, 96-102. https://doi.org/10.1016/j.jad.2014.10.010

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting Depression and Mental Illness on Social Media: An Integrative Review. *Current Opinion in Behavioral Sciences*, *18*, 43-49. https://doi.org/10.1016/j.cobeha.2017.07.005

Han, S., Mao, R., & Cambria, E. (2022). *Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings*. https://arxiv.org/abs/2209.07494

Harrigian, K., Aguirre, C., & Dredze, M. (2020). *Do Models of Mental Health Based on Social Media Data Generalize?* Findings of the Association for Computational Linguistics: EMNLP 2020, https://doi.org/10.18653/v1/2020.findings-emnlp.337

Ji, S. (2022). Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks. *Neural Computing and Applications*, *34*(13), 10309-10319. https://doi.org/10.1007/s00521-021-06208-y

Kim, J., Ma, S. P., Chen, M. L., Galatzer-Levy, I. R., Torous, J., van Roessel, P. J., & Chen, J. H. (2025). *Optimizing large language models for detecting symptoms of comorbid depression or anxiety in chronic diseases: Insights from patient messages*.

LeMoult, J., & Gotlib, I. H. (2019). Depression: A Cognitive Perspective. *Clinical psychology review*, *69*, 51-66. https://doi.org/10.1016/j.cpr.2018.06.008

Losada, D. E., & Crestani, F. (2016). A Test Collection for Research on Depression and Language Use. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 28-39). Springer International Publishing. https://doi.org/10.1007/978-3-319-44564-9_3

Mao, L., He, R., Zhang, S., Ge, Y., Xu, E., Chen, Y., Cong, G., Miao, H., Jiang, Y., & Zhu, H. (2026). Impact of the Interaction Between Screen Time and Activity Interests on Adolescent Depression Risk: Construction of a Predictive Model Based on Machine Learning. *Actas espanolas de psiquiatria*, *54*(2), 500-515. https://doi.org/10.62641/aep.v54i2.2068

Marriwala, N., & Chaudhary, D. (2023). A Hybrid Model for Depression Detection Employing Deep Learning. *Measurement: Sensors*, *25*, 100587. https://doi.org/10.1016/j.measen.2022.100587

Nusrat, M. O., Shahzad, W., & Jamal, S. A. (2024). *Multi Class Depression Detection Through Tweets Employing Artificial Intelligence*. https://arxiv.org/abs/2404.13104

Poświata, R., & Perełkiewicz, M. (2022). *OPI@ LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text Employing RoBERTa Pre-Trained Language Models* Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, https://doi.org/10.18653/v1/2022.ltedi-1.40

Qudar, M. M. A., & Mago, V. (2020). *TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis*. https://arxiv.org/abs/2010.11091

Reavley, N., Jorm, A. F., Carbone, S., Tsiamis, E., & Morgan, A. J. (2025). Testing the Diagnostic Expansion Hypothesis With a Population-Based Survey of Attitudes to Depression in

KMAN-CPN KMAN-Counseling & Psychology Nexus E-ISSN: 3041-9026

Australia. *BMJ Public Health*, *3*(2), e003040. https://doi.org/10.1136/bmjph-2025-003040

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E. D., Kurakin, A., & Li, C. L. (2020). *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence* Advances in Neural Information Processing Systems,

Tharaud, J. B., & Nikolas, M. A. (2025). Emotion Regulation as a Transdiagnostic Link between ADHD and Depression Symptoms: Evidence from a Network Analysis of Youth in the ABCD Study. *Child and adolescent psychiatry and mental health*, *19*, 113. https://doi.org/10.1186/s13034-025-00966-6

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural Language Processing Applied to Mental Illness Detection: A Narrative Review. *NPJ Digital Medicine*, *5*(1), 1-13. https://doi.org/10.1038/s41746-022-00589-7

E-ISSN: 3041-9026