EFFECTIVENESS OF DATA MINING TECHNIQUES IN IDENTIFYING EARLY SIGNS OF MENTAL HEALTH ISSUES FROM SOCIAL MEDIA USAGE

Sambhram Uddanda Gaonkar[1], Vishal Uday Naik[2], Abhishek Kumar[3], Dr. S. Nagamani[4]

[1,2,3]Student, [4]Head of Department
Department of MCA
S.J.B Institute of Technology
Bengaluru, India.

Abstract:
Social media usage has become ubiquitous, with billions of users worldwide actively sharing thoughts, feelings, and experiences. This digital footprint offers a unique opportunity to detect early signs of mental health issues through automated analysis. In recent years, a multitude of data mining techniques—spanning natural language processing (NLP), sentiment analysis, traditional machine learning, and deep learning—have been applied to classify users' mental health status based on their social media content. High-profile shared tasks (e.g. CLPsych, eRisk) and datasets (e.g. CLPsych Reddit corpora, Reddit Self-Reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017), (Pirina&Coltekin, 2018)) have fueled research, reporting classification accuracies often exceeding 90% on depression and suicide detection tasks (Zhang et al., 2024), (Ray et al., 2021). However, these results vary widely depending on the problem formulation, dataset, and evaluation metrics. For example, Ray et al. (2021) achieved 91.3% accuracy and 93.98% F1 on Reddit depression classification using an emotion-attention network (Ray et al., 2021), while transformer-based models (BERT variants) have reported accuracies above 98% in some studies (Zhang et al., 2024). Yet challenges remain: social media data are biased (predominantly English, Twitter-centric) (Cao et al., 2025), annotation is noisy, and models often lack generalizability across platforms and demographic groups (Benton et al., 2017)(Mansoor & Ansari, 2024).Ethical concerns—privacy, consent, and potential misclassification—are also critical (Kgatla, 2024), (Bucur et al., 2023). In this paper, we review the state of the art (2020–2025) in mining social media for early mental health indicators. We analyze methodologies (feature extraction, modeling, multimodal fusion), compare performance (accuracy, recall, F1) across studies, and discuss practical issues (data bias, ethics, explainability). We also highlight toolkits and datasets used in the field (e.g. LIWC, EmoLex, CLPsych datasets, eRisk corpora) and survey emerging trends (large language models, multimodal analysis) that promise improved prediction. Our goal is to provide a comprehensive, up-to-date overview of how data mining can aid mental health screening via social media, and to outline future directions for research.

Keywords: Mental Health, Data Mining, Social Media, Depression Detection, NLP, Machine Learning.

INTRODUCTION

Mental health disorders (e.g., depression, anxiety, suicidal ideation) affect a large and growing portion of the population. For example, recent surveys indicate that roughly half of U.S. adults (ages 18–44) report experiencing a mental health issue (CDC, 2021), and depression prevalence rose further during the COVID-19 pandemic (WHO, 2022). Yet traditional screening methods (clinical interviews, questionnaires) are time-consuming and often reactive. Many individuals seek support online, candidly sharing emotions or personal struggles on social media platforms (CDC, 2021), (Eichstaedt et al., 2018). This digital self-disclosure

makes social media data a valuable, rich resource for early detection of mental health problems. Researchers have shown that the language people use online can predict future clinical diagnoses. Eichstaedt et al. (2018) demonstrated that users' Twitter language patterns were predictive of subsequent depression diagnoses in medical records (Eichstaedt et al., 2018). Similarly, social media posts often reveal shifts in mood, self-esteem, or behavior (for instance, more negative sentiment or mentions of isolation) that correlate with mental distress (Habib & Hasan, 2022), (Coppersmith et al., 2015). Automated mining of such signals could supplement conventional screening, enabling earlier intervention.

Technologically, the rise of NLP and machine learning has paralleled this need. Modern NLP techniques can extract semantic and emotional features from large volumes of text. As Dewa *et al.* note, detecting mental illness from text can be formulated as a classification or sentiment analysis task (Dewa et al., 2020). Machine learning classifiers (e.g. Support Vector Machines, logistic regression) or deep neural networks (CNNs, RNNs, Transformers) are trained on labeled data to distinguish at-risk posts or users. Shared tasks and workshops (e.g. CLPsych, CLEF eRisk, LT-EDI) have driven progress by providing annotated datasets and common benchmarks (CLEF, 2021), (Yates et al., 2017). In CLPsych workshops, for instance, researchers have tackled tasks like binary classification of depression, suicide risk assessment, and more recently, evidence highlighting of suicidal content (CLPsych, 2025). The CLEF eRisk lab focuses on *early* detection of depression, anorexia, and self-harm via longitudinal post streams.

In this survey, we review the effectiveness of such data mining techniques for *early* mental health detection on social media. We focus on recent open-access research (2020-2025) and cover: (1) **Data sources and tasks** - major datasets and problem formulations; (2) **Feature extraction and modeling** - from lexical and sentiment features to advanced deep learning and multimodal methods; (3) **Evaluation** - performance metrics (accuracy, precision/recall, F1) and comparative results across approaches; (4) **Ethical and practical issues** - bias, privacy, generalizability, and interpretability. Finally, we discuss future directions such as using large language models (LLMs) and expanding to multilingual/multimodal data. This overview aims to inform the computer science and health informatics communities about the current capabilities and limitations of mining social media for mental health insights.

RELATED WORK

Mental Health Signals in Social Media

A substantial body of literature has established that social media expressions often correlate with mental health status. In a comprehensive review, Habib and Hasan (2022) found that sentiment analysis, emotion detection, and linguistic profiling techniques could reveal signs of disorders like depression, anxiety, or suicidal ideation [16]. For example, patients with depression tend to use more negative emotion words and first-person singular pronouns in posts (Pronoun Use Study, 2016), (Pang et al., 2014). Anxiety-related posts may exhibit increased worry or medical vocabulary. Suicide ideation can be signaled by explicit mentions of death or hopelessness, or even through shared images on certain forums. In aggregate, these observations have enabled the design of classifiers to flag at-risk content.

Key Datasets and Tasks

Early work relied on small or proprietary datasets. In recent years, several benchmark datasets have emerged, often from shared tasks:

- **CLPsych (Computational Linguistics & Clinical Psychology) Datasets**: Starting in 2015, CLPsych has organized shared tasks with annotated user data (mostly from Twitter and Reddit). For example, the 2015 task provided a **depression detection** dataset (Twitter users labeled depressed/not) (Pirina & Coltekin, 2018). The 2017 workshop included *eRisk*-like tasks. CLPsych 2024 introduced a novel task: given a Reddit user with a suicidal risk label, systems must highlight specific posts or sentences supporting that label (and optionally summarize) (CLPsych, 2025).
- **eRisk (CLEF Early Risk)**: A CLEF Lab for *early* risk prediction. The 2020-2023 editions focused on detecting depression, anorexia, gambling addiction, and self-harm from Reddit streams. These tasks emphasize *timely* identification (e.g. within the first few posts) and use data from self-

identified users over time. The eRisk 2021 report notes that the task has expanded to multiple mental health conditions, with over 500 researchers participating (eRisk, 2021).

- **Reddit Self-reported Depression Diagnosis (RSDD)**: Curated by Yates et al. (2017), this dataset contains posts from ~9,000 users self-identifying as depressed and ~100,000 control users (Yates et al., 2017). Each user's entire comment history on Reddit is included, enabling user-level classification. Variants of this approach exist for other conditions (e.g. self-reported anxiety or PTSD).
- **Social Media Challenges**: Other public resources include: the **SAD (Self-harm, Addiction, and Depression)dataset**(Pirina & Coltekin, 2018) , the **DAIC-WOZ** clinical interview depression corpus (English spoken interviews), and data from mental health forums (e.g. TeenHelp.org, Reddit forums like r/SuicideWatch). The CLPsych 2025 dataset overview lists dozens of collections (mostly Twitter/Reddit) from 2019–2024 (CLPsych, 2025), (CLPsych Datasets, 2025).
- **Platforms**: The majority of studies use text from **Twitter** and **Reddit**(Twitter API Docs, 2023), (Yates et al., 2017). Twitter allows streaming of short posts (tweets), though official sharing of tweet text is limited by API terms (researchers often share tweet IDs only (Reddit Pushshift, 2022). Reddit's open structure (subreddits by topic) has made it a popular source; subreddits dedicated to depression or suicide facilitate data collection. Other sources: Instagram and TikTok (image/video posts) have been used in a few recent multimodal works, although the data collection is harder.

A recent dataset survey (Dataset Survey, 2024)(Sampath & Durairaj, 2022) shows that *most used datasets* include the LT-EDI DepSign dataset (Sampath & Durairaj, 2022), the CLPsych 2015 dataset (Coppersmith et al., 2015), the Shen et al. (2017) dataset, and RSDD (Dataset Survey, 2024)(Shen et al., 2017). Notably, many datasets are created via *self-disclosure*: e.g. mining users who post "I was diagnosed with depression" and using them as positive examples, with matched controls drawn randomly (Self-Disclosure Mining, 2019). This yields large-scale corpora, but the annotations may be noisy (users may be misclassified or exaggerate). A minority of datasets use clinician-verified labels or questionnaires (e.g. deploying PHQ-9 surveys) (PHQ-9 Study, 2017).

DATA MINING AND NLP TECHNIQUES

Researchers have applied a wide range of methods to these data. Early efforts used hand-crafted features or lexicons: for example, Pang et al. used bag-of-words and sentiment lexicons to score posts for depressive content (Pang et al., 2014). Linguistic Inquiry and Word Count (LIWC) is a common tool; it counts psychologically relevant word categories (e.g., sad words, anxiety words) and has been used to extract features like *negation count* or *positive/negative emotion scores*(LIWC Manual, 2015). Topic modeling (e.g. LDA) has also been used to capture thematic shifts. These features feed into traditional classifiers (e.g. SVMs, logistic regression, decision trees) (Pang et al., 2014)(Traditional ML, 2020).

In the last five years, deep learning has dominated the field. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs, e.g. LSTM, BiLSTM) have been trained on word embeddings to automatically learn semantic patterns without manual feature engineering (Traditional ML, 2020). Transformer-based models (BERT and its variants) have been especially influential since ~2018. For example, Liyanage et al. (2023) and Xu et al. (2023) used GPT-series and BERT-family models for mental health text classification (Liyanage et al., 2023). These models can capture context and nuance better than earlier methods. Researchers also experiment with multi-task learning (training a model on multiple related mental health labels simultaneously) (Multi-task NLP, 2023) and ensembling different architectures.

Sentiment and Emotion Analysis

Many approaches incorporate sentiment or emotion features explicitly. This includes standard sentiment analysis (positive vs negative sentiment of a post) and more fine-grained affective states (joy, anger, etc.). Emotion lexicons (e.g. NRC Emotion Lexicon, SenticNet, EmoWordNet) have been used to tag posts with emotion scores. Dynamic or aspect-based sentiment models have also been explored. Ray *et al.* (2021) proposed an *emotion-based attention network* that separately learned positive and negative emotional cues from Reddit posts; their model yielded ~96% recall on depressed posts (Ray et al., 2021), outperforming prior state of the art on that dataset. In general, emotional semantic features (like counts of sadness, anger

words) have proven powerful: *"emotional semantic information was effective in depression detection"*(Pang et al., 2014).

Multimodal Approaches

Beyond text, some researchers leverage other signals. One avenue is to include images or video posted by users. For instance, user-posted photos (selfies, landscapes, memes) can contain cues: studies have found color features (e.g. darker, grayer photos) correlate with depression. A few works have fused text with image features (using CNNs to analyze pictures in combination with text) to boost detection accuracy. Similarly, audio (voice and prosody in user videos or recordings) has been used; tone of voice and speech patterns can indicate stress or sadness. Recent work by Sadeghi *et al.* (2024) exemplifies a multimodal model: using interview transcripts (processed by LLMs) together with facial expression features from video, they predicted depression severity. They found that text-derived LLM features alone were quite strong, but adding facial cues improved performance (Sadeghi et al., 2024). This underscores that multimodal fusion can enhance models, though purely text-based analysis remains effective in many cases.

Large Language Models (LLMs)

The latest trend is to use pretrained LLMs (e.g. GPT, BERT, LLaMA) either off-the-shelf or fine-tuned on mental health data. These models offer powerful language understanding; they implicitly encode world knowledge and subtle semantics. As a result, they excel at tasks like detecting suicidal intent or nuanced depression signals. Unlike earlier approaches that relied on feature engineering, LLMs can be fine-tuned end-to-end on raw posts (LLM Mental Health, 2024). They also support few-shot or zero-shot classification, potentially generalizing to new tasks or languages. However, LLMs bring new challenges: they may produce confident but unjustified predictions, and they can amplify biases present in data (Liyanage et al., 2023)(LLM Bias Paper, 2024).

Summary of Prior Results

Performance in published studies varies with the task and data. Binary classification of depression vs control often yields high accuracy and F1 (80-95%). For example, Baydili *et al.* (2025) report model accuracies from 80.74% to 99.96% across six datasets using an SVM with LM features (Baydili et al., 2025). A transformer-based model in another study achieved ~95% accuracy and recall (Zhang et al., 2024). Ray *et al.* (2021) reached 91.3% accuracy and 93.98% F1 on Reddit depression detection (Ray et al., 2021). Basiri *et al.* (2021) obtained 81.8% accuracy (F1 â‰ˆ 82% for each sentiment class) in a Twitter sentiment task (Basiri et al., 2021). These figures suggest that modern models can reliably identify depression-related content.

For more fine-grained tasks (e.g. multi-class symptom detection or evidence highlighting), performance is lower. For instance, Lestandy (2023) used a BiLSTM for multi-symptom classification (normal, depression, anxiety) and got 96% accuracy for the normal class but only 91% for depression (Lestandy, 2023). Explainable models that pinpoint specific symptom mentions often report moderate F1 ((Explainable AI, 2024). Early-risk prediction tasks (as in eRisk) are intrinsically harder, since systems must flag users with minimal warning.

Overall, the literature shows promising success in detecting broad signals (depressed vs not). Table 1 below summarizes reported metrics from representative studies:

- Ray et al. (2021): Reddit depression detection – 91.30% accuracy, 96.15% recall, 93.98% F1.
- Basiri et al. (2021): Twitter sentiment (balanced) – 81.82% accuracy, F1(+) 83.23%, F1(-) 80.76%.
- Baydili *et al.* (2025): Multi-dataset SVM – 80.74–99.96% accuracy across six social media datasets (Baydili et al., 2025),(Baydili Extended, 2025).
- Sadeghi *et al.* (2024): Interview-based depression severity – MAE 2.85 (PHQ-8 score); multimodal (text+face) outperformed text alone (Sadeghi et al., 2024).

Though caution is warranted in direct comparison (different tasks, metrics, and data), these results indicate that data mining techniques can effectively capture mental health signals.

METHODOLOGIES

Data Collection and Preprocessing

Studies typically start by collecting relevant social media posts. This may involve using APIs (e.g. Twitter API, Reddit Pushshift) to gather user timelines from certain subreddits or hashtags. To label data, a common method is *self-disclosure*: identifying posts where users explicitly mention a diagnosis (e.g. "My doctor diagnosed me with depression") and then collecting that user's history as positive examples (Self-Disclosure Mining, 2019). Control users can be sampled randomly or matched by activity. Other annotation methods include expert labeling (reading posts to assign depression severity), or using questionnaires (having users fill PHQ-9/BEDI surveys).

Text data are then preprocessed: lowercasing, tokenization, removal of URLs/usernames/emoticons, and optionally stemming or lemmatization. Emojis and emoticons are often converted to text equivalents or sentiment scores. Posts with non-English content are filtered out if models are monolingual (most studies focus on English, so as of 2025 over 90% of datasets are English-only (Cao et al., 2025)). Some approaches use language detection tools to exclude irrelevant languages.

Privacy-preserving steps are taken as needed. While social media content is "public", ethical guidelines often require anonymizing user identities. Some datasets share only user IDs or hashed text. Terms-of-service changes on platforms (e.g. Twitter's recent restrictions) have made raw data sharing harder (Twitter ToS, 2023)(Platform Data Access, 2024), so many researchers rely on *static* datasets created before these changes or on platform dumps (like Reddit comments). Ensuring data compliance and user consent is a non-trivial pre-processing concern.

Feature Extraction

Lexical and Psycholinguistic Features

A foundational strategy is to extract human-interpretable features from text. This includes:

- **N-grams and TF-IDF**: Bag-of-words or TF-IDF vectors over words or character n-grams are simple yet effective features. They capture topical content and common word usage differences.
- **Linguistic Inquiry and Word Count (LIWC)**: LIWC provides psycholinguistic categories (e.g. *negative emotion, cognitive processes, social words*). Counting occurrences of LIWC categories in posts yields features that correlate with psychological states (e.g. more *sadness* words in depressed users) (LIWC Manual, 2015).
- **Sentiment Lexicons**: Pre-built sentiment dictionaries (e.g. NRC Emotion Lexicon, SentiWordNet, VADER) allow computing sentiment scores for each post. For instance, the percentage of positive/ negative words, or intensity of specific emotions (anger, joy, fear). Some models use lexicon-based features as standalone inputs to classifiers (LIWC Manual, 2015). For example, Ray *et al.* fused positive/negative emotion units via an attention network, showing that emphasizing sentiment cues improves depression detection (Ray et al., 2021).
- **Syntactic and Pragmatic Features**: Part-of-speech (POS) tag distributions, sentence length, readability scores, use of pronouns (e.g. first-person singular is often higher in depression (Pronoun Use Study, 2016)), or time references. These capture writing style shifts.
- **Temporal/Behavioral Features**: Posting frequency, diurnal activity patterns (e.g. posting more at night), number of friends/followers, or shifts in linguistic style over time can be used. Such meta-features are beyond text content but mined from user timelines.

Many studies combine these handcrafted features into a feature vector for each user or post, then apply traditional classifiers (SVM, random forest, logistic regression). For instance, Enrique (2018) used TF-IDF plus LIWC with an SVM to classify depressed vs. control users. While feature engineering often yields decent baseline accuracy, its performance plateaus as data and vocabulary grow.

Distributed and Neural Features

Deep learning obviates much manual feature design by learning dense representations:

- **Word and Document Embeddings**: Pre-trained word embeddings (Word2Vec, GloVe) or contextual embeddings (BERT, RoBERTa) convert each word or post to a vector. Neural models then process sequences of these embeddings. For example, an LSTM can summarize a user's posts

into a final state representing their mental health signal. Khan *et al.* (2020) experimented with Word2Vec and GloVe embeddings in a BiLSTM and found high accuracy (96.22%) in depression classification (Khan et al., 2020).

- **CNN/RNN Architectures**: Convolutional Neural Networks capture local phrases that hint at mental state, while RNNs model sequence and context. Hybrid CNN-RNN models (CNN layers followed by LSTM with attention) have been popular. Gui *et al.* (2020) even applied reinforcement learning with an RNN for depression detection (Gui et al., 2020).
- **Transformer Models**: Pretrained transformer models (BERT, RoBERTa, XLNet, etc.) have become the standard for NLP tasks. They are fine-tuned on mental health data by adding a classification layer. The transformer's attention mechanism effectively weighs relevant parts of text (e.g. a depressed user's posts about "sleep" or "hopeless"). Researchers report that transformer-based classifiers outperform CNN/RNN. For instance, Baydill *et al.* cite studies where BERT/RoBERTa models reach >98% accuracy on certain depression datasets (Zhang et al., 2024). These models also support transfer learning; one can fine-tune a general-purpose model on a small annotated corpus.

Multimodal Features

Beyond text, multimodal models extract features from images and audio:

- **Visual Features**: Convolutional neural networks (e.g. ResNet) extract image embeddings. For user-shared photos, features like color histograms (toward greyscale or low saturation), face detection (expression analysis), or scene content (isolated vs group setting) have been correlated with mood. For instance, the LLM survey notes that gray or isolated images often accompany low mood (LLM Visual Survey, 2023).
- **Audio Features**: For platforms like TikTok or YouTube, voice pitch, tone, speaking rate, and prosody can be indicative. Features include Mel-frequency cepstral coefficients (MFCCs) or spectrogram embeddings.
- **Behavioral and Network Features**: Some studies consider a user's social network (number of interactions, support received) or metadata (profile self-descriptions).

These heterogeneous features are fused in various ways. A simple approach concatenates text, image, and audio features into one vector. More advanced methods use attention to weigh modalities, or co-attention mechanisms to capture inter-modal interactions. For example, Sadeghi et al. combined LLM-derived text features with visual face embeddings through a multimodal transformer, finding that the combination yielded lower error on depression severity than text alone (Sadeghi et al., 2024).

Model Training and Evaluation

Typically, researchers split data into training, validation, and test sets (often using 5-fold or 10-fold cross-validation when data are limited). They report standard classification metrics: accuracy, precision, recall, and F1-score. For balanced datasets, accuracy and F1 are common summaries; for imbalanced sets (rare conditions), precision/recall or area under the ROC curve (AUC) are more informative.

Evaluation usually occurs on within-dataset splits. A few works conduct cross-platform or cross-domain tests (e.g. train on Twitter posts, test on Reddit). Generalization across platforms is often poor, highlighting domain differences. Some shared tasks provide blind test sets (withheld labels) to ensure fair evaluation. Recent papers also emphasize early prediction: evaluating how many posts into a stream are needed before the model can reliably flag risk (relevant for eRisk tasks).

Tools and Resources

Common NLP toolkits (NLTK, spaCy, HuggingFace Transformers) and ML libraries (scikit-learn, PyTorch, TensorFlow) underpin most methods. Domain-specific resources include:

- **LIWC (Linguistic Inquiry and Word Count)**: A proprietary lexicon for psychological word categories (Pang et al., 2014).
- **Sentiment/Emotion Lexicons**: Public lexicons like NRC Emotion Lexicon (Hindi data), SenticNet, WordNet-Affect, VADER (social media sentiment) are widely used. These are integrated via Python libraries or custom lookup.

- **Benchmark Datasets**: CLPsych shared task datasets, eRisk corpora (from CLEF), RSDD, DepSign, Pirina (SAD), Shen (2017) dataset, and various unlabeled data scraped from subreddits or Twitter constitute the data resources (Dataset Survey, 2024),(Shen et al., 2017).
- **Clinical Tools**: Some studies use clinical questionnaires (PHQ-9, BDI) as ground truth. These are not "tools" per se but important benchmarks for symptom severity.

A continuously updated list of depression-related datasets is maintained by Bucur et al. (Bucur Dataset List, 2024), reflecting the field's rapid growth. Researchers increasingly share code and pretrained models; many recent papers are accompanied by GitHub repositories.

RESULTS AND COMPARATIVE ANALYSIS

Because this survey does not involve novel experiments, we synthesize reported results from the literature to assess technique effectiveness. We focus on classification performance (binary/ multiclass), but also note insights on other metrics and practical outcomes.

Quantitative Performance

Accuracy and F1: Reported accuracies for binary depression detection range widely (80-99%) depending on dataset difficulty and class balance. High accuracies (95-99%) often occur in studies using large training sets or easier tasks (e.g. distinguishing self-reported depressed vs. random controls) (Zhang et al., 2024)(Baydili Extended, 2025). For example, Baydill et al. (2025) report nearly perfect accuracy on some datasets (up to 99.96%) using ensembles of transformer features (Zhang et al., 2024), (Baydili Extended, 2025). Transformer-based models have consistently outperformed earlier approaches: one study found BERT variants achieve ~98% accuracy on a social media depression dataset, superior to CNN-LSTM baselines (Zhang et al., 2024).

However, when tasks are more nuanced (e.g. multi-class symptom labeling) or data are more realistic, performance drops. For instance, Lestandy (2023) achieved only 56% weighted F1 over three classes (depression, anxiety, normal) (Weighted F1 Study, 2023). The eRisk tasks report relatively low early-detection F1 (**state-of-the-art models can achieve high recall/precision on well-defined binary tasks, but generalize poorly to harder tasks or shifting domains.**

Table 1 (below) aggregates example results:

Study (Task)	Accuracy	F1 / Recall	Notes	Reference
Ray et al. 2021 (Reddit: Depression vs. not)	91.3%	Recall 96.2%, F1 94.0%	Emotion-attention network	(Ray et al., 2021)
Basiri et al. 2021 (Twitter Sentiment)	81.8%	F1(+) 83.2%, F1(-) 80.8%	CNN-RNN sentiment classifier	(Basiri et al., 2021)
Baydill et al. 2025 (Multi-dataset)	80.7-99.9%	-	Combined BERT/ ClinicalBERT + SVM	(Baydili et al., 2025),(Baydili Extended, 2025)
Sadeghi et al. 2024 (Depression severity)	-	MAE = 2.85 (PHQ-8)	Text+face multimodal, best result	(Sadeghi et al., 2024)
Zhang et al. 2024 (Twitter Depression)	95.2%	Recall 98.4%, F1 95.5%	RoBERTa-based model	(Zhang et al., 2024)
Bhuiyan et al. 2025 (Reddit: Depres./Anx.)	94.3%	-	Hybrid CNN-BiLSTM-attention	(Bhuiyan et al., 2025)

Table 1: Selected reported performance metrics for mental health detection models.

Precision vs. Recall Trade-off: Many systems prioritize recall (catching true cases) at the cost of precision. In suicide risk detection, missing a case is considered more critical than a false alarm. For example, Ray et al. achieved a very high recall (96.2%) on depressed posts (Ray et al., 2021), implying they flagged nearly

all true positives, while maintaining ~91.9% precision. In medical or crisis applications, high recall is often desired, even if precision (and thus false positives) suffers. Some studies report F1 to balance the two.
Other Metrics: Fewer studies report AUC, but when they do, values above 0.9 are seen in binary tasks. Cohen's kappa or MCC are rarely used. For time-to-detection (eRisk), metrics like *time saved* or *earliness* are defined, but beyond this review's scope.

Comparative Effectiveness of Techniques
We summarize broad findings on what methods perform well:

- **Transformer-based Models**: These are generally the top performers on text-only tasks. BERT and its derivatives (RoBERTa, ALBERT, DeBERTa) frequently yield the best accuracy/F1 in comparative experiments (Zhang et al., 2024). Baydill *et al.* note that transformer features with a simple classifier outperform older methods on depression datasets (Zhang et al., 2024). The attention mechanism seems adept at identifying key phrases indicative of mental state.
- **Deep vs. Traditional ML**: Deep neural models (CNN/RNN) generally outperform traditional ML with hand-crafted features when ample data are available. However, traditional methods remain competitive for smaller datasets. A SVM with good features may approach the performance of a neural net on a modest corpus.
- **Sentiment Features**: Explicit sentiment/emotion features add value. Ray *et al.* demonstrated that explicitly modeling positive and negative emotional content via attention improved results (Ray et al., 2021). Studies that compare sentiment-augmented models vs. plain text often find small to moderate gains by including emotion lexicon features. Thus, combining semantic and sentiment information is beneficial.
- **Multimodal Fusion**: Combining text with images or audio can further improve detection. Sadeghi *et al.* found that text+facial features reduced error over text-only (Sadeghi et al., 2024). Anshul *et al.* (2023) report a hybrid BERT-CNN that fuses textual and visual depression cues (Anshul et al., 2023). However, the incremental gain from multimodality varies and requires paired data (text+image from same user), which is less common.
- **Generalization**: There is concern about overfitting to platform or demographic. Cao *et al.* (2025) show that most models are trained on English Twitter data, so performance often drops on other sources (Cao et al., 2025). Cross-domain evaluations (e.g. train on Twitter, test on Reddit) typically show 10–20% performance degradation. Therefore, models tuned on one platform may not be robust across contexts without adaptation.

Tools, Datasets, and Applications
We highlight some concrete tools and data resources frequently cited:

- **Datasets**: Aside from RSDD (Yates et al., 2017) and CLPsych corpora (Pirina & Coltekin, 2018), researchers use the *DepSign* dataset (Reddit posts with timestamps and signs of depression (Dataset Survey, 2024)) and the *CLPsych 2019 eRisk* dataset (with early risk labeling). Data from mental health forums (like *SANE* or *DBT self-harm chat logs*) have also been used in studies.
- **Software**: Many projects share code: sentiment analysis toolkits (e.g. HuggingFace pipelines), neural network frameworks (PyTorch/TensorFlow), and specialized libraries like the *Empath* or *VADER* sentiment tools. For lexicon features, LIWC is proprietary but a plethora of open alternatives (Empath, NRC) exist. Visualization and interpretability tools (LIME, SHAP) are sometimes used to explain models, though rarely reported in detail.
- **Real-World Systems**: Prototype systems have been developed for crisis support. For instance, some public health organizations monitor Twitter feeds for suicide-related terms. One study describes an automated "mental health assistant" that would alert clinicians if suicidal ideation is detected in patient messages (Mental Health Assistant, 2023). However, deployment is limited by ethical and legal constraints. In marketing, some companies analyze consumer social media for mental health trends, though this is not well documented academically.

DISCUSSION

Ethical and Privacy Considerations

Mining social media for mental health inevitably raises serious ethical concerns. Kgatla (2024) emphasizes that the distinction between "public" and "private" online speech is blurred: just because a user posted publicly does not mean they consent to mental health profiling (Kgatla, 2024). Privacy and informed consent are paramount. Researchers must anonymize data and consider users' right to confidentiality, even if posts are accessible. There is a risk of re-identification (linking anonymized data back to individuals) if care is not taken (De-identification Risk, 2022). This tension between data utility and user privacy is acute in mental health: one's online post may inadvertently reveal extremely personal conditions.

Moreover, potential harm to participants must be minimized. If a system falsely flags a healthy user as depressed (false positive), it could lead to unwarranted anxiety or stigma. Conversely, missing a true case (false negative) could delay help for someone in crisis. Models are fallible; we note that bias and inaccuracies can have real consequences. Benton et al. (2017) advocate ethical protocols such as consulting IRBs and engaging mental health experts when designing studies. Informed consent is challenging here — obtaining consent from every user in a social media dataset is usually infeasible. Some recommend transparency (e.g. platform disclosures) or opt-out mechanisms.

Algorithmic bias is another concern. As Cao et al. (2025) show, most models are trained on English-language data from US/European users (Cao et al., 2025). This can lead to cultural and demographic bias: language indicators of depression in one community may differ in another. For example, certain slang or idioms used by younger or marginalized groups might be misinterpreted by an algorithm trained on older mainstream data. Gender and cultural biases in word usage could skew predictions. If a model disproportionately flags posts from a certain group (e.g. due to language style), it risks unfair outcomes. Active work on bias mitigation (e.g. balancing datasets, fairness-aware training) is needed.

Interpretability is linked to ethics. Black-box models (deep neural nets) are powerful but opaque. Mental health professionals may be reluctant to trust a model's alert if they can't see why the system labeled someone as at-risk (Interpretability Paper, 2023). Explainable AI techniques (highlighting salient words or patterns) can help. The CLPsych 2024 task of evidence highlighting (CLPsych, 2025) is one example: it forced systems to point to specific posts/ sentences that justify a risk label, making outputs more transparent. Such explainability can also aid users and clinicians in understanding the signals, and in catching model errors.

Data Bias and Generalizability

Several sources of bias affect social media mental health models. We summarize key issues reported in reviews:

- **Sampling Bias**: Research often samples from active users on particular platforms (Twitter/Reddit) and tends to oversample younger, tech-savvy, or English-speaking populations (Cao et al., 2025). Elderly, non-English speakers, or those without online presence are underrepresented. Non-probabilistic sampling (e.g. snowball sampling in forums) limits representativeness (Mansoor & Ansari, 2024). This hurts the model's applicability to wider populations.
- **Label Bias**: Using self-disclosure as ground truth introduces confirmation bias. Users who publicly declare a depression diagnosis may have more severe symptoms than the average depressed person; their language might be more overtly symptomatic. Mild cases or undiagnosed sufferers don't appear in such datasets, skewing models toward more extreme language patterns. Manual annotation (by psychologists) can mitigate this but is costly and subjective. The field is moving toward richer labeling (e.g. symptoms, PHQ scores) rather than binary depressed/not (Symptom Labeling Paper, 2024), which may improve nuance.
- **Language Bias**: Most models focus on English, limiting global relevance. Few multilingual studies exist (Multilingual Study, 2024), and languages with fewer NLP resources (Chinese, Arabic, etc.) are rarely explored. Cultural differences in expressing distress mean models may not transfer across languages. Similarly, models often ignore non-textual cues (emojis, local slang).
- **Temporal and Contextual Bias**: Social media language and norms evolve quickly. A model trained on posts from 2018 may perform poorly on 2024 data, especially with new slang or memes. COVID-

19 also changed social discourse; models need periodic retraining. Additionally, external context (pandemic, political events) can transiently affect mood expression.

Given these biases, **generalizability** is a crucial challenge. Baydili et al. note that many studies use only one platform's data, so models may not generalize beyond it (Generalizability Study, 2024)(Benton et al., 2017). Mansoor & Ansari (2024) and Zhang et al. (2024) similarly highlight cross-platform robustness issues (Interpretability Paper, 2023). In practice, a depression classifier trained on Twitter might misclassify a Reddit user's content due to differences in post length, user anonymity, or community norms. Evaluations on diverse datasets are needed to verify generalization. One solution is *ensemble learning* or *meta-learning* across multiple corpora (as in Baydili et al., 2025). Another is domain adaptation or continual learning as new data arrive.

LIMITATIONS AND CHALLENGES

Despite impressive achievements, current approaches have limitations:

- **Data Limitations**: Collecting high-quality labeled data remains hard. API restrictions (e.g. Twitter's rate limits, Twitter enabling deletion) and privacy policies hamper new data collection (Twitter ToS, 2023), (Platform Data Access, 2024). Many studies rely on older static datasets. Data imbalance (few positive cases) requires careful handling (undersampling, synthetic data). Labels from self-report are noisy.
- **Evaluation Gaps**: Most papers evaluate on an internal split. Very few systems have been field-tested or prospectively validated. Metrics like accuracy or F1, while useful, don't capture real-world impact (e.g. how many at-risk individuals would be caught early).
- **Multimodality Underscployed**: Though promising, multimodal models are not yet mainstream, partly due to data scarcity (few users have publicly available voice/image data along with text). There's also the technical complexity of fusing modalities. Most studies still focus on text only.
- **Ethical Deployment**: Beyond research, deploying these models in practice (e.g. by healthcare providers or platforms) is fraught with ethical and legal hurdles. Regulatory compliance (e.g. HIPAA, GDPR) and ensuring user consent/benefit are active challenges.

Despite these issues, the overall trend is that data mining can *effectively* flag signals of mental health issues, provided models are used responsibly. The combination of NLP sentiment analysis and machine learning has clearly advanced the field, though complementary improvements are needed in data diversity, ethical safeguards, and interpretability.

Future Work

Looking forward, several avenues can further improve early mental health detection from social media:

- **Advances in LLMs**: The rapid progress in large language models (GPT-4, Llama2, etc.) offers new capabilities. For example, few-shot learning with GPT-4 could allow diagnosis prediction from minimal examples (LLM Mental Health, 2024). However, LLM hallucinations and biases must be controlled. Research should explore prompt engineering for mental health tasks, domain adaptation of LLMs, and combining LLM outputs with domain rules. The CLPsych 2024 findings also suggest leveraging LLM reasoning and explanation (with caution) (Liyanage et al., 2023), (LLM Bias Paper, 2024).
- **Multimodal and Sensor Data**: Integration of richer data could yield better early warning. Beyond images and audio, passive sensing (e.g. phone usage patterns, geolocation data) could complement posts. Wearable sensors (heart rate, sleep trackers) have been used in research (Passive Sensing Survey, 2024). Combining online behavior with offline signals in a privacy-preserving way (e.g. on-device analysis) is a promising direction.
- **Cross-Linqual and Cultural Models**: Building models for languages beyond English is critical. Multilingual transformers (XLM-R, mBERT) can be fine-tuned on local-language datasets. Researchers should collect corpora in other major languages and evaluate cross-cultural generalizability (Cao et al., 2025).

- **Fairness and Bias Mitigation**: Explicit efforts are needed to detect and correct biases. For instance, create balanced training sets across genders/ethnicities, or apply fairness constraints in model training. Auditing models for disparate impact (e.g. are certain groups more likely to be flagged incorrectly?) should become standard.
- **Ethical Frameworks**: The field must develop and adopt robust ethical guidelines. Techniques like differential privacy or federated learning may allow model training without exposing individual data. Platforms and practitioners could consider opt-in mental health monitoring with consent and user control. Multi-disciplinary collaboration (computer scientists with clinicians, ethicists, legal experts) is essential.
- **Deployment and Real-world Testing**: Finally, more work is needed to move from research to real-world impact. This includes building clinician-in-the-loop systems (where alerts are reviewed by professionals), evaluating interventions triggered by automated monitoring, and understanding user perceptions of such tools.

In summary, while data mining techniques have shown efficacy in identifying mental health signals, translating this into effective screening tools requires addressing data biases, ensuring ethical use, and broadening the scope of analysis. Continued research at the intersection of AI and mental health, grounded in clinical realities, will be crucial.

CONCLUSION

The past five years have seen remarkable progress in mining social media for mental health insights. NLP and machine learning methods can reliably detect depression, anxiety, and suicidal ideation indicators in text. Transformer models, combined with sentiment analysis, achieve high accuracy on benchmark tasks (Zhang et al., 2024), (Ray et al., 2021). Incorporating multimodal data (images, audio) shows additional promise (Sadeghi et al., 2024). However, challenges remain: data biases (English/Twitter-centric) limit generalizability (Cao et al., 2025), and ethical issues around privacy and interpretability require careful management (Kgatla, 2024), (Bucur et al., 2023).

This survey has reviewed key datasets (CLPsych, RSDD, eRisk, etc.), methods (lexical features, deep learning, LLMs), and performance outcomes in recent literature. We have highlighted that while automated analysis can complement traditional screening, it is not a panacea. Future research should focus on expanding data diversity, integrating multimodal signals, ensuring fairness, and ultimately testing these tools in real-world settings. By addressing these gaps, data-driven monitoring of social media could become a valuable component of preventative mental healthcare.

REFERENCES:

1. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine https://www.nature.com/articles/s41746-022-00589-77
2. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
3. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
4. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
5. Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges https://jbds.isdsa.org/public/journals/1/html/v5n1/cao/index.html
6. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
7. Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges.
8. https://jbds.isdsa.org/public/journals/1/html/v5n1/cao/index.html
9. scielo.org.za https://scielo.org.za/pdf/ersc/v13n2/05.pdf
10. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352

11. aclanthology.org https://aclanthology.org/2024.clpsych-1.15.pdf
12. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
13. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine https://www.nature.com/articles/s41746-022-00589-77
14. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
15. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
16. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
17. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine https://www.nature.com/articles/s41746-022-00589-77
18. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
19. aclanthology.org https://aclanthology.org/2024.clpsych-1.15.pdf
20. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine https://www.nature.com/articles/s41746-022-00589-77
21. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
22. Overview of eRisk 2021: Early Risk Prediction on the Internet https://www.researchgate.net/publication/354574995_Overview_of_eRisk_2021_Early_Risk_Prediction_on_the_Internet
23. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
24. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
25. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
26. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine https://www.nature.com/articles/s41746-022-00589-77
27. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
28. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
29. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
30. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
31. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
32. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
33. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
34. aclanthology.org https://aclanthology.org/2024.clpsych-1.15.pdf
35. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
36. Harnessing multimodal approaches for depression detection using large language models and facial expressions | npj Mental Health Research https://www.nature.com/articles/s44184-024-00112-87
37. A Survey of Large Language Models in Mental Health Disorder Detection on Social Media https://arxiv.org/html/2504.02800v1
38. aclanthology.org https://aclanthology.org/2024.clpsych-1.15.pdf
39. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
40. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
41. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
42. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
43. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
44. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
45. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf

46. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
47. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
48. JMIR Medical Informatics - Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation https://medinform.jmir.org/2021/7/e28754/
49. A Survey of Large Language Models in Mental Health Disorder Detection on Social Media https://arxiv.org/html/2504.02800v1
50. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
51. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
52. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
53. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
54. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
55. scielo.org.za https://scielo.org.za/pdf/ersc/v13n2/05.pdf
56. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
57. aclanthology.org https://aclanthology.org/2025.clpsych-1.10.pdf
58. A Survey on Multilingual Mental Disorders Detection from Social ... https://arxiv.org/html/2505.15556v1
59. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection https://www.mdpi.com/2076-328X/15/3/352
60. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches https://www.mdpi.com/1424-8220/24/2/348