The International Journal of Indian Psychology
ISSN 2348-5396 (Online) | ISSN: 2349-3429 (Print)
Volume 13, Issue 4, October- December, 2025
DIP: 18.01.170.20251304, DOI: 10.25215/1304.170
https://www.ijip.in

Review Paper

Social Media Use and Mental Health: Insights from a Targeted Thematic Review

Javed Aslam[1]*

ABSTRACT

This thematic review examines eighteen studies published between 2017 and 2025. It focuses on using social media data to detect and predict mental health conditions, particularly depression and related disorders. The review synthesizes current trends in computational methods, psychological and demographic factors, and ethical issues across major platforms such as Facebook, Instagram, Reddit, and Twitter. The findings show that advanced machine learning models, especially transformer-based approaches, improve the accuracy of mental health detection. Psychological mediators such as social comparison, body image, and loneliness impact the link between social media use and mental health outcomes, especially for adolescents and females. Ethical concerns related to privacy, informed consent, and bias reporting vary greatly among studies, stressing the need for standardized guidelines. While there is promise in this area but challenges remain. These include a lack of diversity in methods, difficulties in generalizing findings beyond Western, English-speaking populations, and integrating results into clinical practice. This review serves as a valuable reference for researchers, clinicians, and policymakers, helping them guide future efforts and ensure the responsible use of social media analytics in mental health surveillance.

Keywords: *Social media, Mental Health, Depression Detection, Machine Learning, Deep Learning, Natural Language Processing, Thematic Review, Psychological Mediators, Demographic Moderators, Ethical Considerations, Computational Psychiatry, Platform-Specific Analysis, Model Performance Metrics*

The rapid growth of social media platforms has greatly influenced mental health research. It has made it possible to collect large amounts of behavioral and language data from people all over the world. Signs of psychological distress such as depression, anxiety, stress, and related disorders, are increasingly visible in users' online activities. This visibility is driving the creation of automated models for predicting mental health on platforms like Twitter, Instagram, Facebook, and Reddit. Even though primary care settings have made progress in diagnosing mental illnesses, a significant number of cases still go undetected or are not treated properly. Social media analytics are starting to help close this gap. They provide ways to identify at-risk individuals quickly, adding to traditional screening methods with new, scalable solutions. (Guntuku et al., 2017; Reece & Danforth, 2017; Arif et al., 2024; Shah et al., 2024; Chancellor & De Choudhury, 2020)

[1]Assistant Professor, Department of Commerce, Bankim Sardar College, West Bengal, India
*Corresponding Author

Received: November 27, 2025; Revision Received: December 01, 2025; Accepted: December 05, 2025

Among the different mental health challenges, depression is the most common focus in computational psychiatry studies. This is mainly due to its widespread occurrence, impact on daily life, and urgent need for early detection. Adolescents and young adults are especially vulnerable because of their specific developmental situations, high use of digital platforms, and ongoing psychological growth. Meta-analytic studies show that problematic and compulsive social media use in these age groups is moderately linked to symptoms of depression, anxiety, and stress. In addition, problematic engagement marked by obsession and difficulty withdrawing seems to have a greater impact than just how often or how long they use social media. (Arif et al., 2024; Shah et al., 2024; Nusrat, Shahzad, & Jamal, 2024; Shannon et al., 2022; Guntuku et al., 2017)

Platforms differ significantly in terms of risk and protective factors. For example, Instagram's focus on images can lead to social comparison, which in turn increases worries about body image, and heightens the risk of low self-esteem and eating disorders among young users. On the other hand, these platforms can also provide spaces for social support, community building, and public health efforts. This is seen in campaigns for suicide prevention, body positivity, and mental health awareness. This dual nature with both positive and negative aspects, highlights the need for a careful and evidence-based evaluation of social media's effects on mental health. (Reece & Danforth, 2017; Shah et al., 2024; Shannon et al., 2022)

Methodologically, the field has quickly changed. Early studies often focused on simple keyword analysis or self-report surveys. Today, more advanced machine learning (ML), deep learning (DL), and natural language processing (NLP) models allow us to detect and predict mental health status from unstructured text, visual, and behavioral data. Transformer-based language models, ensemble methods, and multimodal data fusion show further improvements. They offer the chance for rich, dynamic, and accurate mental health assessments. However, questions about construct validity or how well computational measures reflect clinically important concepts remain unresolved. This is especially true given biases in sampling, annotation, and platform representation. (Chancellor & De Choudhury, 2020; William & Suhartono, 2021; Guntuku et al., 2017; Reece & Danforth, 2017)

These technical and practical innovations raise tough ethical questions. The risks of privacy violations, data misuse, unwarranted surveillance, and social stigma are significant. They require strong data governance practices, informed consent protocols, and transparent reporting standards. A major gap still exists in translating digital phenotyping in clinical settings. Rigorous validation, diverse representation, and collaboration among stakeholders will be essential for responsible innovation. (Shannon et al., 2022; William & Suhartono, 2021; Guntuku et al., 2017; Reece & Danforth, 2017)

Objectives and Scope

This thematic review looks at findings from eighteen diverse studies, covering various regions, populations, and methods. It has three main goals. First, it seeks to outline current trends and methods in assessing mental health through social media. Second, it evaluates the quality and relevance of existing models for detecting and predicting conditions like depression. Third, it highlights the psychological, demographic, platform-specific, and ethical factors that influence these relationships. This review aims to serve as a helpful reference for researchers, clinicians, and policymakers who are exploring the connection between digital technology and mental health. (Shah et al., 2024; Nusrat, Shahzad, & Jamal,

2024; William & Suhartono, 2021; Guntuku et al., 2017; Reece & Danforth, 2017; Arif et al., 2024; Chancellor & De Choudhury, 2020; Shannon et al., 2022)

METHODOLOGY

This review uses a systematic method based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and key standards for thematic and integrative synthesis. A multi-stage process ensured thorough identification, evaluation, and synthesis of studies focused on the detection, prediction, and understanding of depression and related disorders on social media platforms. (Nusrat, Shahzad, & Jamal, 2024; Reece & Danforth, 2017; Shah et al., 2024)

Literature Search and Study Screening

The literature search was conducted mainly through freely accessible platforms, including Google Scholar, ResearchGate, Semantic Scholar, and full-text journal websites. It focused primarily on publications from 2017 to 2025, while also considering foundational papers published shortly before this period. The search used terms like "depression," "anxiety," "mental health," "social media," "machine learning," "deep learning," "NLP," "prediction," and specific platforms such as Twitter, Instagram, Facebook, and Reddit. We also reviewed the reference lists of selected articles and relevant reviews to ensure thorough coverage. (Reece & Danforth, 2017; Shah et al., 2024; Nusrat, Shahzad, & Jamal, 2024)

Inclusion criteria were: (1) empirical studies analyzing original social media data or platforms focused on mental health detection or prediction, and/or review studies providing systematic, thematic, or integrative synthesis of such empirical research; (2) focus on mental health detection or prediction (especially depression, anxiety, stress, suicidality, body image); (3) use of quantitative, qualitative, or mixed methodologies including computational, clinical, or survey-based validation; and (4) publication in peer-reviewed venues in English. Exclusion criteria comprised studies not centered on mental health, lacking a social media focus, being intervention-only, or appearing in non-scholarly literature. (Guntuku et al., 2017; Shah et al., 2024; Nusrat, Shahzad, & Jamal, 2024)

A comprehensive literature search was conducted across multiple databases using predefined keywords related to social media and mental health detection. Studies were systematically screened based on inclusion and exclusion criteria at the title, abstract, and full-text levels. A total of 18 studies comprising 10 empirical and 8 review papers were included for in-depth thematic synthesis.

Data Extraction and Thematic Synthesis

A standardized data extraction process was employed to systematically collect key information from the selected studies, including:

- Study identifiers such as authors, year, and platform
- Available sample characteristics, such as sample size and context, with synthesized insights on demographic vulnerabilities identified in the literature.
- Targeted mental health outcomes and measurement approaches, including validated scales, self-report, and clinical interviews.
- Types of social media data examined across the reviewed studies, including textual content, images, forum discussions, and behavioral activity logs.
- Reported computational and analytical methods, such as machine learning, deep learning, transfer learning, natural language processing, and topic modeling were catalogued and synthesized to identify methodological trends.

- Feature categories reported in the reviewed studies, including textual, visual, behavioral, network, and emotional markers.
- Psychological mediators and demographic moderators investigated (social comparison, body image, loneliness, gender, age)
- Model validation metrics (accuracy, AUC, F1-score, cross-validation)
- Reporting on ethical standards and bias mitigation practices

Findings were synthesized and organized thematically across key domains corresponding to the nine tables: study characteristics, platform-disorder relationships, detection methodologies and performance, psychological and demographic factors, ethical considerations, research trends, and identified gaps. Quantitative results such as AUC and F1-scores were extracted from included studies and descriptively summarized to compare model performance across approaches. Due to methodological heterogeneity and diverse reporting metrics, formal meta-analytic statistical synthesis was not conducted.

RESULTS

Study Characteristics and Overview

Table 1. Comparative Summary of Selected Papers on AI in Mental Health Prediction

Sl No.	Author(s) & Year	Platform(s)	Target Condition	Sample / Population	AI Model(s)	Key Features/ Variables	Performance/Outcome
1	Reece & Danforth (2017)	Instagram	Depression	166 users	Bayesian logistic regression classifiers	Filter use, image patterns	F1 score = 0.647 Significant predictive accuracy Or Effective depression prediction
2	Ameer et al. (2020)	Reddit	Depression, Anxiety, Bipolar disorder, ADHD, & PTSD	16,930 posts	ML, DL, & TL multi-class models	Text Analysis / Language embeddings	RoBERTa Accuracy=0.83, F1-score= 0.83
3	Nusrat, Shahzad, & Jamal, 2024	Twitter	Depression (Bipolar, Major, Psychotic, Atypical, Postpartum)	14,317 Tweets	ML, DL, BERT, Lexicon	Sentiment analysis	BERT Accuracy= 0.96
4	Guntuku et	Facebo	Stress	601	Ngrams,	Languag	Pearson

	al., 2019	ok & Twitter	level	Social media users	LIWC, Elastic-net	e Embeddi ng	correlation r=0.2 6 for user-level stress prediction; TCA improved county-level prediction to r=0.27
5	Turcan & McKeown, 2019	Reddit	Stress	190k Posts	Logistic regressi on/ SVMs, Naïve Bayes, LSTMs, CNNs, BERT embeddi ng	Word2V ec Embeddi ng	F1-Score = 79.8
6	Triantafyllo poulos et al., 2023	Reddit	Depress ion	RSDD Dataset = 28 million posts, Prinia Dateset = 1841 posts	BERT, BiGRU Network	Pretraine d emotion detector	F1-Score RSDD Dataset= 95.65%, Prinia Dateset=6.73%
7	Chandraseka ran et al., 2024	Twitter	Depress ion	229 depress ed users, 2,46,63 7 Tweets	CorEx topic modelin g, SVM, Naïve Bayes, Logistic Regressi on	Topic Modelin g	F1-Score SVM= 91.21 LR= 90.58 NB= 79.67
8	Murarka et al., 2020	Reddit	Mental Illness: Depress ion, Anxiety , Bipolar disorder , ADHD, PTSD	17,159 Posts	RoBER Ta, LTSM, BERT	Text Labellin g	RoBERTa F1-Score= 0.89 (Posts + titles)
9	Eichstaedt et al., 2018	Facebo ok	Depress ion	683 Patients	LDA, Logistic regressi on	Predictio n model	AUC = 0.69 (before diagnosis), F1 score = 0.66
1 0	Liu et al., 2022	Facebo ok	Lonelin ess, Depress	34,59,8 54 posts	BERT embeddi ngs,	Text classific ation	BERT only: r = 0.201

			ion	from 2,986 users	LDA topic modelin g	and Labellin g	

Table 1 summarizes ten empirical studies that used artificial intelligence and machine learning models to predict or detect mental health conditions from social media data. The studies cover various platforms, including Instagram, Reddit, Twitter, and Facebook, and target different mental health conditions such as depression, anxiety, bipolar disorder, stress, and PTSD.

Key Thematic Findings

A. Platform Diversity and Mental Health Conditions

The studies show that different social media platforms provide valuable data for mental health prediction. Research based on Instagram focused on visual features, like filter use and image patterns. In contrast, text-heavy platforms such as Reddit, Twitter, and Facebook allowed for linguistic analysis. The mental health conditions studied ranged from specific diagnoses, like depression and bipolar disorder, to broader issues, such as stress and loneliness. This shows the wide applicability of AI-driven analysis.

B. AI Model Sophistication

The research shows a trend toward more complex models. It progressed from traditional machine learning methods, such as logistic regression and Support Vector Machines, to more advanced deep learning architectures. Transfer Learning models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT) proved to be especially effective. For example, RoBERTa achieved an F1-score of 0.83 (Ameer et al., 2020), while BERT reached 96% accuracy in classifying depression (Nusrat, Shahzad, & Jamal, 2024). These transformer models outperformed standard machine learning classifiers by capturing deeper semantic relationships in user-generated text.

C. Performance Metrics and Clinical Relevance

Performance indicators varied significantly across studies. F1-scores ranged from 0.647 to 0.96, with accuracy scores reaching as high as 96%. Notably, Triantafyllopoulos et al. (2023) achieved very high performance with an F1-score of 95.65% on the RSDD dataset. Eichstaedt et al. (2018) also demonstrated predictive validity before clinical diagnosis, with an AUC of 0.69. This suggests potential for early detection. However, the wide variation in performance metrics highlights the importance of model selection and dataset characteristics in ensuring predictive success.

D. Feature Extraction and Data Scale

The studies used various techniques for feature extraction, including linguistic embeddings like Word2Vec and LIWC, sentiment analysis, topic modeling, and image pattern analysis. Sample sizes varied widely, ranging from 166 Instagram users (Reece & Danforth, 2017) to 28 million Reddit posts (Triantafyllopoulos et al., 2023). This illustrates the scalability of computational methods. Larger datasets allowed for more effective model training, as shown by higher performance metrics in studies with millions of posts.

E. Interdisciplinary Integration

The body of work demonstrates a blend of computational linguistics, machine learning, and mental health research. Techniques such as N-grams, LIWC (Linguistic Inquiry and Word Count), topic modeling, and transfer learning constitute advanced methods used to explore mental health expression in digital spaces. This interdisciplinary approach places these studies at the intersection of data science and clinical psychology.

Social Media Platforms and Mental Health Outcomes

Table 2. Social Media Platforms and Disorders

Platform	Main Disorders Studied	Representative Outcomes/Markers	Supporting Studies
Facebook	Depression, stress level, Loneliness	Language, medical record linkage	Guntuku et al. (2017), Eichstaedt et al. (2018), Liu et al., 2022
Instagram	Depression, Body image, Well-being	Filter use, photo features, comparison	Reece & Danforth (2017), Faelens et al. (2021), Balamurugan et al. (2023)
Reddit	Depression, Bipolar disorder, stress, Mental Illness, Suicide risk, Stress, EDs	Forum text, subgroup participation	Shah et al., 2024, Ameer et al. (2020), Turcan & McKeown, 2019, Triantafyllopoulos et al., 2023, Murarka et al., 2020
Twitter	Depression, stress level, Well-being, Bipolar	Tweet text, emotional valence	Nusrat, Shahzad, & Jamal, 2024, Guntuku et al., 2019, Chandrasekaran et al., 2024,

Table 2 summarizes the mental health disorders studied across four major social media platforms: Facebook, Instagram, Reddit, and Twitter. It includes key outcome markers and representative studies that investigated these relationships.

Platform-Specific Mental Health Focus

A. Facebook as a Depression and Stress Indicator

Facebook allows for self-disclosure and social networking. This platform has become useful for detecting depression, stress levels, and loneliness. Research based on Facebook data mainly uses linguistic markers derived from user posts and comments. It has added advantage of potential medical record linkage for validation purposes. Studies by Guntuku et al. (2017), Eichstaedt et al. (2018), and Liu et al. (2022) demonstrated that language patterns on Facebook, such as negative sentiment and first-person singular pronouns, and absolutist language are closely linked to self-reported depression and stress levels. This makes Facebook a good option for large-scale mental health monitoring.

B. Instagram: Visual and Comparative Mental Health Markers

Instagram's visual-centric nature makes it suitable for studying depression, body image concerns, and overall well-being. Unlike text-based platforms, Instagram allows the analysis of visual aspects like filter use, photo aesthetic patterns, color saturation, and exposure of faces. Notably, Reece & Danforth (2017) found that certain image traits, such as cooler colors, increased saturation, and lower brightness, were predictive of depression. The

platform's tendency for comparison, where users showcase idealized version of thier lives, leads to body image dissatisfaction and lower well-being as documented by Faelens et al. (2021) and Balamurugan et al. (2023).

C. Reddit: The Most Diverse Mental Health Research Platform

Reddit stands out as the platform with the widest range of mental health issues studied, including depression, bipolar disorder, stress, general mental illness, suicide risk, and eating disorders (EDs). This variety comes from Reddit's structure: thematic subreddits, like r/depression and r/bipolar, attract users actively discussing or seeking support for specific conditions. Researchers have used forum text and patterns of subgroup participation as primary markers. The large amount of available data and explicit condition-related communities have allowed studies by Shah et al. (2024), Ameer et al. (2020), Turcan & McKeown (2019), Triantafyllopoulos et al. (2023), and Murarka et al. (2020) to achieve strong predictive models for various mental health conditions.

D. Twitter: Emotional Expression and Real-Time Mental Health Signals

Twitter's real-time, text-based communication structure allows researchers to capture immediate emotional expressions and provides access to large longitudinal datasets. Research has focused on depression, stress levels, well-being, and bipolar disorder, with tweet text and emotional valence serving as primary analytical features. The short and spontaneous nature of tweets reflects authentic emotional states, making them valuable for detecting transient stress and mood fluctuations. Studies by Nusrat, Shahzad, & Jamal (2024), Guntuku et al. (2019), and Chandrasekaran et al. (2024) have utilized lexicon-based sentiment analysis and transformer models to classify emotional valence and predict mental health conditions.

E. Cross-Platform Insights

The table reveals important cross-platform patterns: depression is the most frequently studied condition across all four platforms, highlighting its prevalence and the availability of related linguistic and behavioral markers in social media. In contrast, some conditions, like eating disorders and suicide risk, are predominantly studied on Reddit, where condition-specific communities provide concentrated data sources. The variety of outcome markers, including linguistic, visual, and behavioral participation patterns, indicates that each platform offers distinct methodological affordances for mental health research. This requires platform-specific analytical approaches while allowing for cross-platform validation studies.

Detection Approaches, Features, and Performance

Table 3. Detection Approaches and Model Performance

Study (Year)	Model/Approach	Features Used	Validation/Metric	Key Result/Outcome
Guntuku et al. (2017)	Ngrams, LIWC, Elastic-net, SVM	Text, behavior	AUC= 0.70 to 0.91	Facebook, Twitter language predicts depression
Reece & Danforth (2017)	Bayesian logistic regression classifiers	Images, text	F1 score = 0.647	Filter/image preferences predict

				symptoms
Ameer et al. (2020)	ML, DL, & TL multi-class models	Text Analysis/Language embeddings	Accuracy=0.83, F1-score= 0.83	Multi-class mental illness detection

Table 3 summarizes the methodological approaches and predictive performance metrics from three important studies, each employing distinct analytical strategies to detect mental health conditions from social media data. The comparison shows the shift from traditional machine learning methods to hybrid approaches that include deep learning and transfer learning, alongside the corresponding improvements in model performance.

Traditional Machine Learning Approaches: Guntuku et al. (2017)

Guntuku et al. (2017) pioneered the use of linguistic and behavioral features extracted from Facebook and Twitter posts to predict depression. They used N-grams (contiguous word sequences capturing local semantic context) and LIWC (Linguistic Inquiry and Word Count), a psychologically validated lexicon that quantifies linguistic dimensions like emotional tone, cognitive processes, and personal references. These features were fed into Elastic-net regularization and Support Vector Machine (SVM) classifiers, achieving an impressive Area Under the Curve (AUC) range of 0.70 to 0.91. This range reflects variation across platforms and population subgroups, with Twitter language showing slightly lower predictive validity than Facebook. This approach proved that text-based linguistic markers are reliable indicators of depression, supporting the core hypothesis that mental health conditions show distinct patterns in social media language.

Multimodal Feature Integration: Reece & Danforth (2017)

Reece & Danforth (2017) took a fundamentally different approach by analyzing Instagram images with text, recognizing that visual self-presentation might convey mental health information. They used Bayesian logistic regression classifiers and showed that they could predict depression symptoms based on image features like filter use (especially cool-toned filters), image saturation, brightness level, and whether human faces were present or absent. Although their F1-score of 0.647 was lower than Guntuku et al.'s results, this innovative multimodal approach highlighted that depression is reflected not only in linguistic expression but also through visual aesthetics choices, opening a new dimension of mental health research on visual platforms.

Advanced Deep Learning and Transfer Learning: Ameer et al. (2020)

Ameer et al. (2020) advanced the field by bringing together several machine learning methods such as traditional machine learning, deep learning, and transfer learning, into multi-class classification models. They used language embeddings (dense vector representations of words that capture semantic meaning) to automatically learn complex linguistic patterns without needing manually designed features like N-grams or LIWC. They achieved 83% accuracy and an F1-score of 0.83, showing that modern deep learning models could classify multiple mental health conditions, a notable improvement over earlier studies that focused only on binary classification (depressed vs. non-depressed). The use of language embeddings and transfer learning marks a significant shift toward end-to-end learning models that learn discriminative patterns directly from raw text without relying on predetermined linguistic dictionaries.

Psychological and Demographic Moderators

Table 4. Psychological Mediators and Moderators

Mediator/Moderator	Direction/Effect	Measured Impact/Outcome	Reference/Study
Social Comparison	Negative	↑ Depression, ↓Self-esteem	Reece & Danforth (2017), Faelens et al. (2021)
Body Image	Negative	↑ Anxiety, ↑Depression	Balamurugan et al. (2023)
Self-esteem	Protective	↑ Self-esteem ↓ Depression	Shannon et al. (2022)
Loneliness	Negative	↑ Anxiety, ↑Depression	Shah et al., 2024, Owens et al. (2024)
Gender/Age	Moderator	↑ Risk adolescent/female	Faelens et al. (2021), Balamurugan et al. (2023)

Table 4 summarizes the psychological mechanisms by which social media use influences mental health outcomes. Social comparison and body image concerns stand out as primary negative mediators, increasing depression and anxiety while lowering self-esteem. In contrast, self-esteem serves as a protective factor, buffering against depressive symptoms. Loneliness demonstrates bidirectional effects; functioning as both a consequence and an exacerbating mediator of social media-related mental health deterioration. Demographic factors, especially gender and age, function as moderators. Adolescent females showing greater vulnerability to negative psychological outcomes. These findings clarify how social media engagement affects mental health, showing both risks and protective psychological processes.

Table 5. Demographic Variability in Impact

Demographics	Major Findings/Impact	Supporting Studies
Adolescents	Most vulnerable for negative effects	Shannon et al. (2023), Balamurugan et al. (2023)
Non-English/cultures	Understudied, risk of bias	Owens et al. (2024)
ages and genders	females and adolescents often at higher risk for negative outcomes	Faelens et al., 2021

Table 5 highlights critical demographic disparities in social media's mental health effects. Adolescents emerge as the most vulnerable group, experiencing severe negative psychological outcomes. There are clear gender differences, with females demonstrating heightened susceptibility to depression, anxiety, and body image concerns compared to males. There is a significant gap in research on non-English-speaking and culturally diverse populations. This limits how generalizable the findings are and introduces possible sampling bias towards Western, English-speaking groups. These demographic variations highlight the need for population-specific research and culturally sensitive interventions to tackle mental health risks related to social media.

Ethics and Societal Implications

Table 6. Ethical Considerations Across Reviewed Papers

Paper	Informed Consent	IRB Mentioned	Demographic Data	Bias Control
Guntuku et al. 2017	No	No	Partial	Partial
Reece & Danforth	Yes	Yes	No	Partial
Chancellor et al.	Yes	Yes	-	Yes
Nusrat, Shahzad, & Jamal, 2024	YES	-	YES	YES
Guntuku et al. 2019	YES	YES	Partial	YES
Eichstaedt et al., 2018	YES	YES	YES	-
Liu et al., 2022	YES	YES	YES	Partial

Table 6 reveals substantial heterogeneity in ethical reporting standards across the reviewed literature. Recent studies from 2018 onward consistently mention informed consent and IRB approval. However, earlier studies from 2017 reveal notable gaps in ethical documentation. Most recent publications include informed consent procedures and institutional review, although reporting of demographic data is still inconsistent. Majority of studies document bias control measures, but several papers do not clearly show how they address potential selection or measurement bias. These variations indicate that ethical standards in computational mental health research are evolving. They also emphasize the need for standardized ethical reporting practices to ensure transparency, reproducibility, and the protection of research participants.

Table 7. Ethical, Legal, and Societal Issues

Study	Main Ethical Issue	Summary/Key Point	Implication/Note
Guntuku et al., 2017	Privacy, Consent	Need for strict de-identification	Fundamental for model deployment
Reece & Danforth (2017)	Data privacy, consent	markers of depression are observable in Instagram user behavior,	Psychological changes appear in social media behavior and can be computationally detected.
Shah et al., 2024	Data availability, diversity in users behavior, Choice of the classification models, Dataset bias	Data Collection biggest challenge, unstructured, irregular data	Roadmap on designing mental illness detection models (preprocessing, feature extraction, and classification techniques)
Chancellor & De Choudhury, 2020	Consent, Lack of standardized methods	Review of 75 studies shows growth in social media–based mental health prediction but with inconsistent methods.	Standardized reporting and collaboration are needed for reliable, ethical social media mental health predictions.
Nusrat, Shahzad, & Jamal, 2024	Consent, dataset limitation, privacy and security of sensitive health data	Data construction was a challenge,	Machine Learning and Deep Learning models successfully predicted depression types
Guntuku et al., 2019	privacy concerns, data protection,	People may use social media less when stressed,	computational methods could provide real-time, individual

	transparency about the indicators derived	potentially reducing the model's accuracy.	feedback from social media data to support simple, technology-assisted stress and mindfulness interventions
Owens et al. (2024)	Privacy, Consent	acquiring sizable data is challenging	emotion conveyed in social media posts is an underrepresented topic
Eichstaedt et al., 2018	privacy, informed consent, data protection, and data ownership	Social media–based depression screening could become more feasible and accurate.	linking mental health diagnoses with social media content can reveal depression-related patterns and may enable scalable early identification of depressed individuals.
Liu et al., 2022	Informed consent,	Loneliness is linked to self-focused thinking and heightened attention to environmental cues.	Interventions should target maladaptive social cognitions and strengthen social relationships.

Table 7 outlines the main ethical concerns identified across the reviewed studies. Privacy and informed consent emerge as universally recognized issues. They require strict de-identification protocols and transparent user notification before data collection and analysis. Data availability and quality create operational challenges, especially with dataset bias and the unstructured nature of social media data. A significant gap is the lack of standardized reporting methods and ethical guidelines across the studies, which calls for a collaborative development of consistent protocols. The studies stress that strong ethical frameworks, covering data protection, informed consent, and transparency regarding algorithmic indicators, are fundamental for responsibly using computational mental health detection models. Moreover, the potential intervention applications and the underrepresentation of emotion-related topics in dataset diversity highlight the need for multidisciplinary ethical oversight from various fields in social media-based mental health research.

Research Trends and Gaps

Table 8. Trends in Methodologies Over Time (2015–2024)

Year Range	Dominant Methodology	Notes
2013–2018	Traditional ML (SVM, Decision Trees, Logistic Regression), Lexicon / Bag-of-Words / TF-IDF	Early text mining methods
2018–2021	Deep Learning (CNN, LSTM)	Shift to neural nets
2021–2024	Transformer-based Models & Multimodal	State-of-the-art, increasing real-world relevance

Table 8 shows how computational methods in social media-based mental health research have evolved over time. The field has clearly evolved from traditional machine learning methods from 2013 to 2018. During this period, researchers used hand-crafted features like lexicon-based methods and bag-of-words representations. From 2018 to 2021, there was a shift toward deep learning models, such as Convolutional Neural Networks and Long Short-Term Memory networks. These models allowed for automatic feature extraction from raw textual and visual data. Since 2021, transformer-based models and multimodal approaches have become dominant. They offer top performance and are more applicable to real-world

situations. This shift in methods shows how the field has matured and improved its ability to capture complex psychological indicators from social media data.

Table 9. Gaps and Limitations in Current Research

Study (Year)	Reported Limitation	Recommended Future Direction
Guntuku et al. (2017)	Limited generalizability	Detect undiagnosed cases
Faelens et al. (2021)	Reliance on self-report, cross-sectional	Longitudinal, clinical validation
Shah et al. (2024)	Annotation bias, privacy concern	enhance sentiment analysis, use advanced DL models, target underserved conditions, and expand geographic coverage.
Owens et al. (2024)	bias in methods,	Scope exists for broader coverage to enable follow-up studies. Future mental health prediction research should identify and report data biases to improve reliability.

Table 9 summarizes the key research limitations and future recommendations identified in the reviewed studies. Generalizability is a major concern, as the field needs to expand beyond mostly Western, English-speaking populations and demographics. Methodological limitations include a heavy dependence on self-reported symptoms and cross-sectional designs, which limit causal conclusions. Longitudinal studies with clinical validation are necessary to establish temporal relationships and predictive validity. Data quality issues, especially annotation bias and uneven geographic representation, compromise model reliability and transferability. Future research must prioritize advanced deep learning models for sentiment analysis, target underserved mental health conditions and populations, broaden geographic and cultural representation, and systematically identify and report data biases. These recommendations highlight that rigorous methodological advancement, combined with transparent bias reporting and longitudinal validation, is essential for translating computational mental health detection from research to clinical practice.

DISCUSSION

Synthesis of Findings

This review synthesizes evidence from 18 studies that collectively demonstrate the evolving capacities and complexities of using social media data to understand and predict mental health challenges. Instead of looking at each platform or study separately, the integrated findings highlight a dynamic interplay between technological advancements, psychological mechanisms, demographic vulnerabilities, and ethical imperatives.

A key point is that social media platforms are not interchangeable data sources; each one has unique features that shape how mental health indicators appear and can be detected. Text-based platforms like Facebook, Twitter, and Reddit primarily facilitate linguistic and sentiment analysis. Reddit's thematic communities providing particularly rich data on diverse mental health conditions. Instagram's visual orientation allows for analysis that includes image features, such as filter usage and color patterns, linking these to outcomes such as body image and depression. These specific characteristics of each platform require

tailored computational methods rather than generic models, showing the need for careful methodology.

The field has witnessed significant algorithmic evolution. Traditional machine learning methods, which heavily depended on hand-engineered text features, have gradually given way to deep learning and current state-of-the-art transformer-based models such as BERT and RoBERTa. These advanced models use complex semantic representations to achieve superior detection accuracy. They show progress in recognizing the subtle language patterns that relate to mental health states. However, performance can differ based on dataset size, condition complexity, and platform. This shows that no single model universally dominates, and context remains critical.

Beyond algorithmic sophistication, psychological processes identified as mediators such as social comparison, body image, self-esteem, and loneliness serve as important links to the findings in established mental health theory. These contructs help explain the ways in which social media engagement can either increase or reduce psychological distress, especially among adolescents and females who tend to be more vulnerable. Demographically, the noticeable lack of representation for non-English-speaking and culturally diverse populations reveals a significant gap that limits the generalizability and ethical strength of current models.

Ethical issues are present in the reviewed literature, revealing both progress and gaps. More recent studies increasingly document informed consent and institutional oversight. However, inconsistencies in bias control, data anonymization, and transparency remain pervasive. These ethical challenges highlight an urgent need for standardized frameworks that ensure respectful, equitable, and privacy-conscious research and clinical practice.

Collectively, the integration of technological advances, psychosocial frameworks, demographic insights, and ethical analysis positions social media–based mental health detection at a crossroads. Innovative algorithms hold promise for scalable early identification and monitoring. However, turning these findings into clinical practice responsibly requires interdisciplinary collaboration. This includes robust longitudinal validation, culturally responsive model development, and sustained ethical vigilance.

Implications for Clinical Practice, Policy, and Future Research

Building on the synthesized findings of this review, this section outlines the implications for clinical practice, policy development, and future research directions to enhance the use and impact of social media-based mental health detection.

1. Clinical Practice Implications

Social media-based mental health detection shows promise as a complementary tool to traditional clinical assessments. Evidence demonstrating predictive validity prior to clinical diagnosis suggests potential for early identification and timely intervention referral. Clinicians should interpret linguistic and behavioral indicators from social media as reflective but not definitive of psychological distress. Computational outputs should inform clinical judgment rather than replace it. Effective integration requires standardized protocols for model validation, continuous performance monitoring, clinician training, and informed patient consent regarding data use. Vulnerability-centered design principles are crucial to prevent stigma and to prioritize patient dignity and autonomy.

2. Policy and Regulatory Recommendations

There is a pressing need for standardized ethical frameworks governing the development and deployment of computational mental health tools. Regulatory policies should mandate informed consent, transparent algorithmic documentation, compliance with data protection standards and independent audits assessing algorithmic bias. Public health investments should prioritize research inclusivity focusing on underserved and culturally diverse populations. International collaboration is vital for harmonizing consistent ethical standards related to data governance and the prevention of exploitative practices. Policies must safeguard individual rights in algorithmic profiling and emphasize that computational tools augment rather than substitute human-centered care.

3. Future Research Directions

Future studies should address several priorities to strengthen the field's evidence base and ethical grounding:

1. Longitudinal validation through prospective designs incorporating clinical outcomes to establish causal inference and predictive robustness.
2. Inclusion of linguistically and culturally diverse populations to improve generalizability and fairness.
3. Rigorous examination of construct validity to ensure computational markers map onto clinically meaningful phenomena, ideally integrating psychological theory with machine learning frameworks.
4. Exploration of advanced multimodal methodologies harnessing combined textual, visual, temporal, and social network features, accompanied by explainable AI techniques to improve model transparency.
5. Development of ethical implementation pathways through multidisciplinary collaboration involving technologists, clinicians, ethicists, and community stakeholders to balance innovation with participant protection.
6. Expansion beyond detection toward intervention design, leveraging real-time social media insights for personalized mental health support and treatment delivery.

CONCLUSION

This thematic review examined the evolving landscape of computational mental health detection using social media data. The 18 studies analyzed show a growing area marked by important technological advances, particularly the adoption of transformer-based models that achieve high accuracy (83–96%) in detecting conditions such as depression, anxiety, bipolar disorder, and related disorders. Each major social media platform, including Facebook, Instagram, Reddit, and Twitter, offers unique analytic opportunities. Approaches designed for specific platforms outperform generic methodologies.

The review shows consistent evidence that psychological distress can be observed in the way people use social media, particularly in their linguistic expression and visual presentation. Important psychological factors like social comparison, body image concerns, and loneliness stand out among vulnerable populations such as adolescents and females. These findings align with established theories on social media's effects on mental health and validate the potential of using computational methods for detection.

Nonetheless, important limitations temper enthusiasm. Methodological heterogeneity, the dominance of Western English-speaking samples, limited longitudinal validation, and inconsistent ethical practices all restrict the generalizability and practical relevance of the result. Significant disparities in ethical reporting and bias reduction highlight the urgent need

for standardized protocols. Additionally, the lack of representation from culturally diverse populations and the scarcity of cross-cultural validation limit the models' applicability across broader contexts.

Successfully translating computational mental health detection into clinical and public health practice necessitates multidisciplinary collaboration. This includes technological refinement, strong ethical oversight, clinical validation, and attention to equity across different populations. Future research must emphasize longitudinal designs, inclusivity, robust construct validation, and identifying systematic biases. It is important to have policy frameworks that enforce ethical standards, ensure strong data governance, and provide regulatory oversight. This can support responsible innovation and reduce the risk of misuse.

In conclusion, computational analysis of social media data shows strong potential for scalable mental health surveillance and early detection. However, realizing this promise requires moving beyond accuracy metrics and prioritizing ethical standards, inclusivity, and meaningful clinical integration. The field is at a pivotal juncture. As technological capabilities grow, they must be accompanied by strong accountability to ensure that innovations empower and protect vulnerable populations. Collaborative efforts involving researchers, clinicians, policymakers, ethicists, and affected communities will be vital to harness this promise while minimizing risks related to bias, privacy, and inequity.

REFERENCES

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

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science, 6*(15), 1-12. https://doi.org/10.1140/epjds/s13688-017-0110-z

Arif, M., Ameer, I., Bölücü, N., Sidorov, G., Gelbukh, A., & Elangovan, V. (2024). Mental illness classification on social media texts using deep learning and transfer learning. *Computacion y Sistemas, 28*(2), 451-464. https://doi.org/10.13053/cys-28-2-4873

Shah, S. M., Aljawarneh, M. M., Saleem, M. A., & Jawarneh, M. S. (2024). Mental illness detection through harvesting social media: A comprehensive literature review. *PeerJ Computer Science, 10*, e2296. https://doi.org/10.7717/peerj-cs.2296

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

Nusrat, M. O., Shahzad, W., & Jamal, S. A. (2024). Multi class depression detection through tweets using artificial intelligence. *arXiv Preprint*, arXiv:2404.13104. https://doi.org /10.48550/arXiv.2404.13104

Shannon, H., Bush, K., Villeneuve, P. J., Hellemans, K. G., & Guimond, S. (2022). Problematic social media use in adolescents and young adults: Systematic review and meta-analysis. *JMIR Mental Health, 9*(4), e33450. https://doi.org/10.2196/33450

William, D., & Suhartono, D. (2021). Text-based depression detection on social media posts: A systematic literature review. *Procedia Computer Science, 179*, 582-589. https://doi.org/10.1016/j.procs.2021.01.043

Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019). Understanding and measuring psychological stress using social media. *Proceedings*

of the International AAAI Conference on Web and Social Media, 13(01), 214-225. https://doi.org/10.1609/icwsm.v13i01.3223

Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. *arXiv Preprint*, arXiv:1911.00133. https://doi.org/10.48550/arXiv.1911.00133

Owen, D., Lynham, A., Smart, S., Pardiñas, A., & Camacho Collados, J. (2024). AI for analyzing mental health disorders among social media users: Quarter-century narrative review of progress and challenges. *Journal of Medical Internet Research, 26*, e59225. https://doi.org/10.2196/59225

Triantafyllopoulos, I., Paraskevopoulos, G., & Potamianos, A. (2023). Depression detection in social media posts using affective and social norm features. *arXiv Preprint*, arXiv:2303.14279. https://doi.org/10.48550/arXiv.2303.14279

Chandrasekaran, R., Kotaki, S., & Nagaraja, A. H. (2024). Detecting and tracking depression through temporal topic modeling of tweets: Insights from a 180-day study. *npj Mental Health Research, 3*(1), 62. https://doi.org/10.1038/s44184-024-00107-5

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2020). Detection and classification of mental illnesses on social media using RoBERTa. *arXiv Preprint*, arXiv:2011.11226. https://doi.org/10.48550/arXiv.2011.11226

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences, 115*(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

Balamurugan, G., & Vijayarani, M. (2025). Filtering reality: Navigating Instagram's influence on adolescent mental health. *Journal of Education and Health Promotion, 14*, 84. https://doi.org/10.4103/jehp.jehp_483_24

Liu, T., Ungar, L. H., Curtis, B., Sherman, G., Yadeta, K., Tay, L., Eichstaedt, J. C., & Guntuku, S. C. (2022). Head versus heart: Social media reveals differential language of loneliness from depression. *npj Mental Health Research, 1*(1), 16. https://doi.org/10.1038/s44184-022-00014-7

Faelens, L., Hoorelbeke, K., Cambier, R., van Put, J., Van de Putte, E., De Raedt, R., & Koster, E. (2021). The relationship between Instagram use and indicators of mental health: A systematic review. *Computers in Human Behavior Reports, 4*, 100121. https://doi.org/10.1016/j.chbr.2021.100121

Acknowledgment

The author(s) appreciates all those who participated in the study and helped to facilitate the research process.

Conflict of Interest

The author(s) declared no conflict of interest.

How to cite this article: Aslam, J. (2025). Social Media Use and Mental Health: Insights from a Targeted Thematic Review. *International Journal of Indian Psychology*, *13*(4)*,* 1856-1873. DIP:18.01.170.20251304, DOI:10.25215/1304.170

LIST OF ABBREVIATIONS

- ML = Machine Learning Classifier/ Machine Learning Models
- DL = Deep Learning
- TL = Transfer Learning

- BERT = Bidirectional Encoder Representations from Transformers
- LIWC = Linguistic Inquiry and word Count
- SVM = Support Vector Machine
- LSTM = Long Short-Term Memory
- CNN = Convolutional Neural Network
- BiGRU = Bidirectional Gated Recurrent Units
- CorEx = Correlation Explanation (in topic modeling)
- RoBERTa = A Robustly Optimized BERT Pretraining Approach
- LDA = Latent Dirichlet Allocation
- TCA = Transfer Component Analysis
- ED = Eating Disorders
- N-grams = Contiguous sequences of N items from a given sequence of text
- AUC = Area Under the Curve
- ↑ = Increase
- ↓ = Decrease
- IRB = Institutional Review Board
- TF-IDF = Term Frequency-Inverse Document Frequency