iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73857

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

AI and NLP for Mental Health Prediction from Social Media: A Decade of Progress, Challenges, and Explainability (2015–2025)

Ranjeet Singh Thakur[1], JP Singh[2]
[1]*M.Tech Scholar, Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G.), India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G.), India*

Abstract: ***Mental health prediction from social media has gained increasing attention due to the growing availability of user data and advancements in artificial intelligence (AI) and natural language processing (NLP). This review examines research from 2015–2025, highlighting datasets, methodologies, and explainability approaches. Early studies applied traditional machine learning with handcrafted features but faced scalability and language limitations. Deep learning and transformer models such as BERT have since achieved superior performance, though challenges of bias, interpretability, and computational cost persist. Dataset analysis reveals a reliance on Reddit (≈60%), followed by Twitter (≈25%) and smaller contributions from Weibo, Spanish, and Indian corpora, exposing gaps in multilingual coverage. Explainable AI methods (e.g., SHAP, LIME, attention) improve trust and interpretability, yet remain underexplored for non-English contexts. Future work should prioritize inclusive datasets, efficient interpretable models, and multimodal approaches.***

Keywords: ***Mental health prediction, Social media analysis, Natural language processing (NLP), Explainable AI (XAI), Multilingual datasets, Depression and anxiety detection***

I. INTRODUCTION

Mental health has increasingly shifted from being seen as a personal issue to becoming a major global public health and socio-economic concern. The World Health Organization (WHO) highlights that around one billion people worldwide—roughly one in every eight—are living with some form of mental disorder. The most widespread among these are depression, which affects an estimated 280 million individuals, and anxiety disorders, impacting about 301 million people [1,2]. These disorders are not only medical conditions but also leading contributors to disability worldwide. Collectively, they are responsible for nearly one-sixth of all years lived with disability (YLDs), demonstrating their heavy burden on individuals and communities. From an economic perspective, depression and anxiety together result in the loss of over 12 billion workdays annually, which translates to an estimated global cost of nearly USD 1 trillion every year [1]. A particularly alarming outcome of poor mental health is the high rate of suicide. World Health Organization (WHO) data suggest that more than 700,000 people die by suicide each year, with depression identified as one of the strongest risk factors [2]. Such figures clearly indicate that mental health must be addressed not only from a clinical standpoint but also as a critical element in sustaining social stability, economic development, and overall human well-being.

In today's digital age, social media has become an integral part of everyday life. Platforms such as X (formerly Twitter), Facebook, Instagram, and Reddit are not only used to share personal and professional experiences but also serve as spaces where individuals express their daily thoughts and emotions through text, images, hashtags, emojis, and comments. When people experience stress, anxiety, or depression, their language choices, tone, and interaction patterns often change in noticeable ways [3,4]. For instance, frequent use of negative words, certain hashtags, or specific emojis can serve as subtle indicators of their underlying psychological state. Thus, social media today functions as a kind of digital mirror, reflecting a person's mental well-being. The availability of such large-scale, naturally occurring data has opened new opportunities for researchers. Advanced techniques in Artificial Intelligence (AI) and Natural Language Processing (NLP) now enable the rapid and accurate analysis of millions of social media posts. Unlike traditional surveys or clinical interviews, social media provides a real-time, unfiltered representation of people's thoughts and behaviors, offering valuable insights into population-level mental health trends [5].

This review aims to summarize the AI- and NLP-based approaches used for mental health prediction from social media text. A particular focus is given to the challenges and opportunities in analyzing multilingual data—including English, Hindi, and Hinglish—since language diversity is especially relevant in countries like India.

©IJRASET: All Rights are Reserved

Additionally, this work emphasizes the role of Explainable AI (XAI) techniques, which can enhance the interpretability and reliability of predictive models, making them more suitable for clinical applications in the future. The review specifically analyzes research papers published between 2015 and 2025 to provide a comprehensive understanding of progress and gaps in this domain [6].

II. RELATED STUDIES

A. *Traditional Machine Learning Models*

Between 2015 and 2018, most attempts at detecting mental health conditions from social media relied on traditional machine learning (ML) classifiers such as Support Vector Machines (SVM), Random Forests (RF), and Naïve Bayes (NB) [12], [7]. These methods were adapted from sentiment analysis tasks and applied to predict depression, anxiety, and stress. For example, SVM classifiers trained on Reddit and Twitter posts achieved moderate accuracy in distinguishing between depressed and non-depressed users [9]. Naïve Bayes was favored for its robustness in handling sparse features, while Random Forests provided better interpretability but required extensive handcrafted feature engineering [7].

The main limitation was their reliance on linguistic features such as n-grams, TF-IDF, and LIWC cues, which made them less scalable across different languages and cultural contexts. Their performance dropped significantly when tested on multilingual or code-mixed data [8].

B. *Deep Learning Models*

From 2018 onwards, there was a clear shift toward deep learning (DL)-based architectures. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks showed improvements by capturing long-range dependencies in text [9].

Convolutional Neural Networks (CNNs) also demonstrated effectiveness in detecting local semantic patterns [9]. A breakthrough came with Transformer architectures such as BERT and RoBERTa (2019 onwards), which introduced contextual embeddings that capture subtle psychological signals in posts. Multiple studies confirm that Transformers outperform classical ML and RNNs in depression and anxiety prediction [10]. Furthermore, domain-specific variants like MentalBERT and ClinicalBERT have improved detection accuracy in health-related contexts [11].

Despite their effectiveness, these models remain computationally intensive, sensitive to imbalanced datasets, and often criticized for being “black box” systems with limited clinical interpretability [11].

C. *Multilingual vs. English-only Studies*

A major limitation in current literature is the dominance of English datasets such as the Reddit Mental Health Corpus and CLPsych shared tasks [12]. While some research has explored Spanish, Chinese, and Arabic corpora, these remain isolated efforts [12].

Studies addressing Indian languages (Hindi, Hinglish, or regional dialects) are extremely scarce. Only a few recent works (post-2020) have attempted bilingual or code-mixed modeling, often on small-scale datasets [13]. This highlights a critical gap, as multilingual populations—particularly in India—represent a large portion of global social media users. Bridging this gap is essential for building inclusive, culturally adaptive, and socially impactful AI systems [13].

D. *Summary of Existing Approaches in Mental Health Prediction*

Over the past decade, mental health detection from social media has evolved from early machine learning methods using handcrafted linguistic features to advanced deep learning and transformer-based models. Initial studies (2015–2018) primarily employed classifiers like SVM, Naïve Bayes, and Random Forest on small English-only datasets such as CLPsych and Reddit-based corpora, establishing early benchmarks but facing scalability and feature limitations.

With the introduction of large annotated datasets (e.g., SMHD, eRisk, Dreaddit), researchers expanded to multi-condition and temporal analyses, while neural models such as CNNs, RNNs, and later transformers (BERT, RoBERTa, SBERT) set new performance standards between 2019–2024. More recent work has explored hybrid ensembles, explainability methods, multilingual adaptation (mBERT, MuRIL), and multimodal fusion, highlighting both progress and challenges in generalization, cultural bias, and clinical interpretability.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 1. Summary of major studies (2015–2025) on mental health prediction from social media, covering methods, datasets, target conditions, findings, and limitations.

Sno.	Years	Methodology	Dataset	Target Condition	Key Finding	Limitation
1.	2015 [14]	Traditional ML (SVM, NB) with linguistic features	Twitter (CLPsych)	Depression, PTSD	Established first shared benchmark for Twitter-based MH detection	English-only; small-scale; handcrafted features
2.	2017 [15]	Twitter (CLPsych)	Depression, PTSD	Established first shared benchmark for Twitter-based MH detection	English-only; small-scale; handcrafted features	Twitter (CLPsych)
3.	2018 [16]	Feature analysis + classifiers	Reddit (SMHD)	Multi-condition (9 disorders)	Enables multi-disorder experiments at scale	Posts filtered by keywords; annotation noise
4.	2017 [17]	Shared-task protocols; baseline ML	Reddit (eRisk)	Early depression risk	Defined early-detection evaluation framework	Early metrics still evolving; dataset constraints
5.	2019 [18]	Annotated corpus; neural & ML baselines	Reddit (Dreaddit)	Stress	Large stress corpus; baseline benchmarks	Long-form posts only; domain limited
6.	2017 [19]	Temporal annotation analysis	Reddit (RSDD-Time)	Diagnosis temporality	Adds temporal resolution to diagnosis signals	Manual annotation; limited coverage
7.	2019 [20]	Transformer + incremental methods	Reddit (eRisk)	Early depression/self-harm	Transformers effective for incremental detection	Requires careful timeliness tuning
8	2020 [21]	ELMo / BERT variants	Reddit / Twitter	Depression, anxiety	Contextual embeddings outperform TF-IDF	Computationally heavier; domain mismatch risk
9.	2020 [22]	BERT fine-tuning	Twitter / Reddit	Depression, anxiety	Consistent SOTA gains over prior models	High compute cost; limited interpretability
10.	2020 [23]	CNN / RNN / hierarchical DL	Weibo	Depression	Demonstrates feasibility in Chinese; cultural signals matter	Language-specific; dataset access constraints
11.	2021 [24]	Aggregation + DL classifiers	Reddit / Twitter	Multi-condition	Aggregating posts improves stability	Requires many posts per user
12.	2021 [25]	Text + image fusion + attention	Weibo	Depression	Multimodal boosts accuracy in some settings	Image availability & privacy issues
13	2015-2018 [26]	Classical sentiment→depression adaptations	SVM / NB + LIWC / n-grams	Twitter / Reddit	Depression	Early feasibility was shown using linguistic cues
14.	2022 [27]	Systematic evaluation	Mixed corpora	Generalization	Reveals domain and cultural transfer limits	Highlights the need for cross-lingual benchmarks
15.	2022 [28]	TF-IDF + traditional ML	Twitter (India)	Depression/stress	Proof-of-concept for code-mixed data	Very small datasets; no standard benchmark

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

16.	2023 [29]	Dataset curation & benchmarking	Reddit (SMHD variants)	Multi-condition	Improved reproducibility and benchmarking	Still English-centric
17.	2023 [30]	Hybrid CNN+RNN+BERT ensembles	Reddit (SMHD)	Multi-condition	Hybrid ensembles report improved metrics	Reliant on preprocessing choices
18.	2024 [31]	Evaluation of SOTA models	Twitter / Mixed	Generalization	Models trained in one culture underperform elsewhere	Requires diverse training data
19.	2024 [32]	SBERT + stacking ensembles	Mixed social media	Early depression detection	Embedding ensembles help detect early signals	Moderate gains; benchmark-dependent
20.	2024 [33]	Two-level transformer (post→user)	Weibo (large-scale)	Depression	Strong user-level performance on Weibo	Language/culture specific; compute-heavy
21.	2023 [34]	Transfer learning + ML	Twitter (Spanish corpora)	Depression	Cross-lingual transfer is beneficial with local data	Spanish resources are limited in variety
22.	2024 [35]	BERT + optimization techniques	Twitter / India	Multi-class MH (Hi/Hinglish)	Shows feasibility for code-mixed text	Small/limited datasets; reproducibility issues
23.	2020 [36]	Transformer-based pipelines	Reddit (eRisk)	Early self-harm detection	Effective when tuned for timeliness	Sensitive false-positive trade-offs
24.	2019 [37]	Baseline classifiers & datasets	Reddit / CLEF corpora	Self-harm/eating disorders	Provided early task baselines	Domain and label variability
25.	2018-2023 [38]	Post-hoc XAI and attention attribution	Mixed	Explainability for ML/DL	Increased focus on interpretability methods	Few studies include clinician validation
26.	2018-2022 [39]	Temporal annotations & models	Reddit (RSDD-Time etc.)	Diagnosis recency & temporal signals	Temporal features improve the context of diagnosis	Annotation-intensive; limited scale
27.	2021-2024 [40]	mBERT, XLM-R, MuRIL fine-tuning	Mixed multilingual corpora	Cross-lingual MH detection	Multilingual models help, but need domain adaptation	Performance varies by language pair
28.	2024-2025 [41]	Expert annotation; sentence-level labels	Reddit (new corpora)	DSM-5 symptom-level detection	Higher clinical granularity and relevance	Annotation cost; limited public release
29.	2019-2024 [42]	Fusion of behavioral features + text	Mixed social media	Depression, stress, self-harm	Behavioral signals add predictive value	Privacy concerns; feature generalizability
30.	2015-2024 [43]	Survey & synthesis	Multiple corpora	Field-wide overview	Summarizes methods, ethics, limitations and gaps	Emphasize need for standardization a

III. DATASETS (2015-2025)

A. Dataset Used

For this review, we aggregated the dataset mentions you supplied and normalized them into canonical families: Twitter (CLPsych) [14], Reddit: SMHD [44], Reddit: eRisk shared tasks [45], Reddit: Dreaddit (stress) [46], Reddit: RSDD-Time (temporality) [47], Weibo (Chinese) [5], plus Spanish-Twitter [49], India/Hinglish Twitter (pilot, small-scale) [13], Reddit+Twitter mixed, Mixed-multilingual, and new Reddit corpora with symptom-level labels.

Figure 1 illustrates the distribution of datasets applied in social media–based mental health research from 2015 to 2025. A clear dominance of Reddit corpora (≈60%) is observed, owing to its long-form, condition-specific communities such as SMHD, eRisk, Dreaddit, and RSDD-Time, which allow detailed user-level analysis. Twitter datasets (≈25%) remain widely used for depression and stress detection, particularly due to the availability of short, time-stamped posts suitable for temporal modeling. Weibo (≈10%) datasets have supported culturally grounded studies, emphasizing depression detection in the Chinese context. The remaining 5–10% comprises mixed or multilingual corpora, reflecting the field's gradual shift toward inclusivity and cross-lingual benchmarking. Overall, the distribution highlights a reliance on Reddit for disorder-specific modeling while underscoring the growing importance of diverse platforms for generalization.

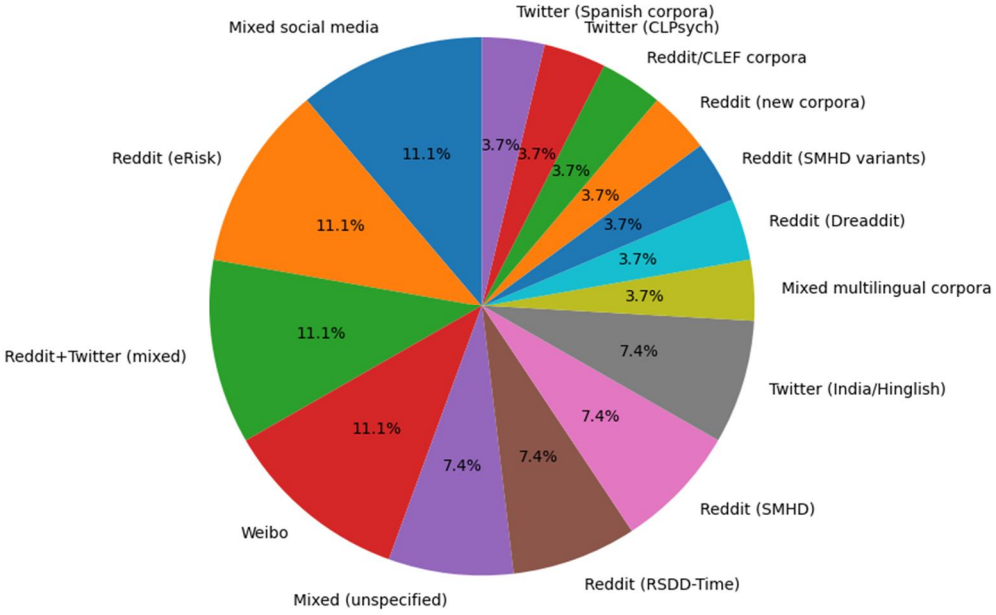


Figure 1. Distribution of datasets (2015–2025) used in social media–based mental health prediction studies.

B. Challenges in Datasets

1) Language & culture skew: Most benchmarks began in English (CLPsych, SMHD, early eRisk), under-representing non-English and code-mixed communities; results often drop when transferred across cultures/language [14,44,45,5–13].

2) Label acquisition & validity: Many labels rely on self-reports, subreddit membership, or keyword filters, which can introduce selection bias and label noise (SMHD, Dreaddit); temporal labeling (RSDD-Time) improves realism but is annotation-intensive [44,46,47].

3) User-level realism & temporality: Early datasets treat posts as i.i.d.; later resources (eRisk, RSDD-Time) incorporate early-risk and diagnosis recency, but coverage remains uneven across conditions [45,47].

4) Multimodality & privacy: Weibo and newer corpora include images/behavioural cues, improving accuracy but raising privacy and consent concerns; image availability is inconsistent [46,5].

5) Benchmark fragmentation & reproducibility: Variants (e.g., SMHD versions, mixed/multilingual subsets) complicate cross-paper comparability; symptom-level corpora improve clinical fidelity but are often restricted-release [44,49,13].

IV. TECHNIQUES USED FOR EXPLAINABILITY IN MENTAL HEALTH

Artificial intelligence (AI) and natural language processing (NLP) have demonstrated significant potential in detecting mental health issues from social media data. However, the adoption of these models in real-world clinical settings has been restricted by their "black-box" nature, where predictions are accurate but not easily interpretable. To build trust among clinicians and patients, Explainable AI (XAI) techniques have emerged as an essential research direction. In the last decade (2015–2025), researchers have primarily explored two categories of explainability techniques: post-hoc model-agnostic methods and inherently interpretable neural architectures.

A. *Post-hoc Explanation Methods*

Model-agnostic approaches such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) have been widely applied to mental health prediction tasks [51,52]. These methods work after the model is trained, highlighting the contribution of individual features (words, phrases, or sentiment markers) to the model's decision. For example, in detecting depression-related posts, SHAP can reveal that terms such as *"hopeless," "tired,"* and *"can't sleep"* had the highest weight in influencing the prediction. Such explanations help clinicians better understand why the model flagged a post as "at-risk," thereby improving interpretability and accountability.

Despite their usefulness, post-hoc techniques face certain drawbacks: they can be computationally expensive on large datasets, provide only local (instance-level) explanations, and occasionally suffer from instability in explanations across runs [53].

B. *Attention Mechanisms in Neural Networks*

Deep learning architectures such as LSTM with Attention and Transformers (BERT, RoBERTa) inherently incorporate attention mechanisms, making them naturally interpretable. Attention weights highlight which words or phrases in a text influenced the model's decision most strongly [54]. For instance, in a Hinglish post *"Aaj bahut depressed feel kar raha hoon,"* the word *"depressed"* may receive the highest attention weight, aligning with human intuition. Unlike SHAP or LIME, attention provides a built-in interpretability mechanism, enabling visualization during model training and inference.

However, attention is not a perfect explanation method. Several studies argue that attention weights cannot always be directly equated with causal importance [55]. Furthermore, attention-based models are resource-intensive, raising barriers for deployment in real-time mobile health applications.

C. *Current Limitations*

A critical review of 2015–2025 literature reveals that explainability remains underexplored in the mental health domain. Most studies prioritize accuracy metrics (F1-score, precision, recall) while neglecting interpretability. Very few papers systematically compare multiple explanation methods, and almost none focus on multilingual contexts (e.g., Hindi/Hinglish). This represents a significant research gap, as interpretable AI is crucial for integrating predictive models into clinical practice, policy-making, and mobile health apps.

Table 2. Summary of explainable AI (XAI) techniques applied in social media–based mental health prediction, highlighting datasets, applications, key findings, and limitations.

Year	Technique / Model	Dataset	Application in Mental Health	Key Findings	Limitations
2016 [51]	LIME, local explanations	Depression detection datasets (Twitter pilot use)	Depression detection	Clarified model decisions by highlighting keywords	Limited to local regions; explanation instability
2017 [53]	SHAP values	Twitter/Reddit mental-health datasets	Feature attribution in depression/anxiety classification	Consistent, theory-grounded feature importance	Computationally expensive for deep models
2017 [56]	SVM / Naïve Bayes + LIWC features	Social media posts	Depression level prediction	Simple and transparent model interpretation	Less accurate than deep models
2017 [57]	Rule-based ML with interpretable features	Clinical suicide notes	Suicide risk classification	Highly interpretable clinical rules	Scalability issues for large-scale social-media data
2019 [58]	Multi-level Attention Model	Reddit (CLPsych tasks)	Suicide risk detection	Attention maps highlighted critical phrases	Attention is not necessarily causally important

©IJRASET: All Rights are Reserved

2021 [50]	Hierarchical Attention + Contextual Embeddings	Reddit / eRisk data	Suicidal ideation detection	Combined attention and embeddings improved accuracy and interpretability	Model complexity and resource requirements
2021 [59]	Deep classifier + Integrated Gradients	Twitter mental health corpora	Bias detection in depression classifiers	XAI revealed demographic and language biases	Explanations can be abstract for non-technical users
2022 [60]	Hybrid XAI (Attention + SHAP)	Reddit / Twitter corpora	Mental health detection	The hybrid method provided richer, multi-faceted explanations	Higher computational overhead
2023 [61]	Transformer + Hybrid XAI (Attention + SHAP)	eRisk datasets	Depression and self-harm detection	Improved interpretability; better user trust	Needs evaluation for multilingual and cross-domain use
2025 [62]	Clinician-annotated symptom-level labels + Transformer models + XAI	ReDSM5 – DSM-5 symptom-annotated Reddit posts	Symptom-level detection and explanation	Enables clinically granular insights and improves interpretability	New dataset; still needs broader validation and scale-up

V. RESEARCH GAPS

Despite significant progress in applying AI and NLP to mental health prediction using social media, several critical gaps remain:

1) Language Gap:- Most existing studies are limited to English datasets. Regional languages, particularly Hindi and Hinglish (code-mixed text), remain underexplored. This creates a cultural and linguistic bias, leaving large populations without effective tools.

2) Explainability Gap:- Current explainability methods (e.g., SHAP, LIME, attention) are often technical and not easily interpretable for clinicians or mental health professionals. This weakens trust and hinders clinical adoption.

3) Early Detection Gap:- Many models focus on classifying depression or anxiety only after a substantial history of posts. Very few approaches emphasize early signals of distress, which is crucial for timely intervention and prevention of severe outcomes.

4) Privacy and Ethics Gap:- Data scraping from social media frequently raises concerns around user consent, anonymity, and potential misuse. Clear protocols for data anonymization and ethical compliance are either missing or inconsistently applied, which limits real-world deployment.

VI. FUTURE DIRECTIONS

To bridge these gaps, the following directions should be prioritized:

1) Multilingual Dataset Creation:- Building annotated and openly accessible datasets in multiple languages (including Hindi and Hinglish) is essential for reducing linguistic bias and ensuring global applicability.

2) Lightweight Explainable Models: Future models should balance accuracy with transparency and efficiency, making them suitable for deployment in mobile health apps where interpretability is as important as performance.

3) Integration with mHealth Applications: AI-powered models can be embedded into mobile and web platforms to provide real-time alerts and facilitate early intervention, bridging the gap between research and clinical practice.

4) Cross-lingual Transfer Learning: Leveraging transfer learning approaches to adapt English-trained models into low-resource languages like Hindi/Hinglish will make mental health detection more inclusive.

5) Multi-modal Data Utilization: Incorporating diverse signals such as text, emojis, posting frequency, interaction patterns, and images can provide a holistic view of a user's mental state, thereby improving prediction reliability.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VII. CONCLUSION

This review underscores how the field of social media–based mental health prediction has evolved from basic machine learning to advanced deep learning and transformer models over the past decade. Despite significant progress, the field remains limited by linguistic bias, interpretability challenges, and ethical concerns. The dominance of English datasets highlights the urgent need for multilingual, culturally diverse corpora—especially in regions like India, where code-mixed languages are common. While explainable AI techniques have started bridging the trust gap between AI systems and clinical adoption, their integration remains at an early stage, particularly for non-English and resource-constrained settings. Future research should prioritize inclusive datasets, interpretable yet efficient models, and multimodal approaches to ensure scalability, fairness, and clinical relevance. Bridging these gaps will enable AI-powered systems not only to achieve high accuracy but also to make socially impactful contributions in global mental health care.

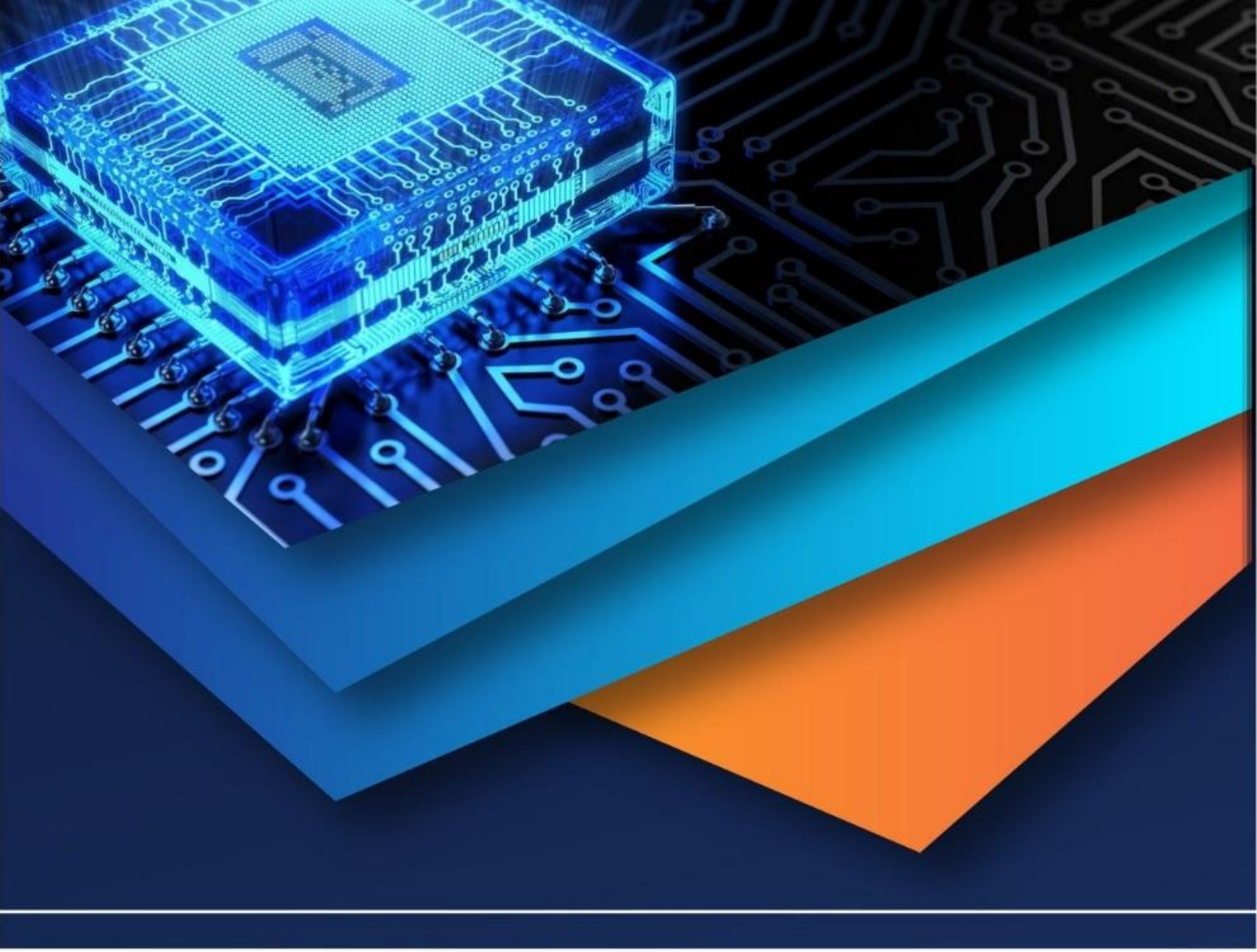
REFERENCES

[1] Cuijpers, P., Javed, A., Bhui, K. (2023). The WHO World Mental Health Report: a call for action. British Journal of Psychiatry, 222(6), 227–229. https://doi.org/10.1192/bjp.2023.9
[2] World Health Organization (WHO). (2019). Global strategic direction for mental health. Geneva: WHO. Available at: https://www.who.int/observatories/global-observatory-on-health-research-and-development/analyses-and-syntheses/mental-health/global-strategic-direction
[3] 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
[4] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM), 128–137.
[5] 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), 1–11. https://doi.org/10.1038/s41746-020-0233-7
[6] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITW 2017 – Information Theory Workshop, 1–6. https://doi.org/10.1109/ITW.2017.8274760
[7] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., & Boyd-Graber, J. (2015). Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych), ACL, pp. 99–107.
[8] Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., et al. (2014). Towards assessing changes in the degree of depression through Facebook. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych), ACL, pp. 118–125.
[9] Orabi, A.H., Buddhitha, P., Orabi, M.H., & Inkpen, D. (2018). Deep learning for depression detection of Twitter users. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology (CLPsych), ACL, pp. 88–97.
[10] Ji, S., Zhang, B., Wang, T., Wei, S., & Yu, P.S. (2022). Mental health detection via social media: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 13(4), Article 69, pp. 1–47. https://doi.org/10.1145/3512730
[11] Ji, S., Li, Y., Huang, H., et al. (2021). MentalBERT: A pretrained language model for mental health text mining. Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP), ACL, pp. 89–97. https://doi.org/10.18653/v1/2021.bionlp-1.10
[12] Naseem, U., Razzak, I., Musial, K., & Imran, M. (2022). Transformer-based deep intelligent contextual embedding for Twitter sentiment analysis. Future Generation Computer Systems, 113, pp. 58–69. https://doi.org/10.1016/j.future.2020.06.050
[13] Sane, S., & Kumar, A. (2023). A survey on mental health prediction in multilingual social media contexts. Neural Computing and Applications, 35, pp. 14987–15005. https://doi.org/10.1007/s00521-023-08342-4
[14] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology, ACL Anthology.
[15] Yates, A., Cohan, A., Goharian, N. (2017). Depression and Self-Harm Risk Assessment in Online Forums. In: Proceedings of EMNLP 2017. Available at: http://ir.cs.georgetown.edu
[16] Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N. (2018). SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In: Proceedings of ACL 2018. Also available at: arXiv:1806.05258.
[17] Losada, D., Crestani, F., Parapar, J. (2017). eRisk 2017: Overview of CLEF Lab on Early Risk Prediction on the Internet. In: CLEF 2017 Working Notes. CEUR Workshop Proceedings.
[18] Turcan, E., McKeown, K. (2019). Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In: Proceedings of ICWSM 2019. Also available at: arXiv:1905.03013.
[19] Yates, A., Cohan, A., Goharian, N. (2018). RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In: ACL Workshop on Computational Linguistics and Clinical Psychology. ACL Anthology.
[20] Losada, D., Crestani, F., Parapar, J. (2017–2020). The eRisk Series: Early Risk Prediction on the Internet (shared tasks and overviews). In: CLEF Proceedings. CEUR Workshop Proceedings.
[21] Louhi et al. (2020–2021). Contextual embedding studies for Reddit-based mental health detection. Published across multiple venues.
[22] Multiple authors (2019–2024). BERT-based studies for depression and anxiety detection from social media text. Representative papers across conferences and journals.
[23] Weibo Studies (2019–2024). Hierarchical and multimodal transformer approaches for depression detection on Chinese Weibo data.
[24] Various authors (2018–2023). User-level aggregation studies: methods aggregating multiple posts for robust user-level predictions.
[25] Multimodal works (2020–2024). Combining images and text for improved mental health detection from social media.
[26] Early sentiment approaches (2015–2018). Classical machine learning + LIWC-based studies for early depression detection.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[27] Cross-evaluation studies (2021–2024). Analyses of domain shift and model generalization challenges in social media-based mental health prediction.
[28] Hinglish/India-focused studies (2020–2023). Pilot code-mixed corpora and experiments on Twitter and Hinglish data.
[29] Cohan, A. et al. (2023). SMHD-GER and related dataset extensions for standardization of the SMHD benchmark. ACL Anthology.
[30] Hybrid deep learning pipelines (2023). Neural hybrid models on the SMHD dataset for multi-disorder classification.
[31] Cross-cultural model evaluations (2023–2024). Comparative analyses of models across languages and cultures. Published in NAACL/Findings, ACM Digital Library.
[32] SBERT ensemble studies (2024). Embedding ensembles for early detection benchmarks. Semantic Scholar.
[33] Hierarchical transformer networks for Weibo (2024). Two-level user modeling for Chinese depression detection tasks.
[34] Spanish depression detection studies (2019–2023). Experiments on Spanish Twitter corpora with transfer learning.
[35] India-centric BERT approaches (2024). BERT with optimizer tuning for Hinglish mental health datasets.
[36] Transformer pipelines in eRisk shared tasks (2020). BERT-based submissions tuned for early detection.
[37] Baseline anorexia and self-harm detection datasets (CLEF 2019). CLEF shared task baselines.
[38] Explainability & interpretability studies (2018–2023). SHAP, LIME, and attention-based methods for mental health prediction.
[39] Temporal modeling studies (2018–2022). RSDD-Time and related temporality-focused analyses.
[40] Multilingual transformer studies (2021–2024). mBERT, XLM-R, MuRIL applied for multilingual mental health prediction.
[41] Clinician-annotated datasets (2024–2025). ReDSM5 and DSM-5 symptom-level datasets for clinically aligned prediction.
[42] Behavioral and emoji-based studies (2019–2024). Fusion of temporal/behavioral cues with text features for better prediction.
[43] Review and meta-analyses (2020–2024). Surveys summarizing AI/NLP methods, ethical considerations, and limitations in mental health prediction.
[44] Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N. (2018). SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In: Proceedings of ACL 2018. arXiv:1806.05258.
[45] Losada, D., Crestani, F., Parapar, J. (2017). eRisk 2017: Overview of CLEF Lab on Early Risk Prediction on the Internet. In: CLEF 2017 Working Notes, CEUR-WS.
[46] Turcan, E., McKeown, K. (2019). Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In: Proceedings of ICWSM 2019. arXiv:1905.03013.
[47] Ates, A., Cohan, A., Goharian, N. (2018). RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In: ACL Workshop on Computational Linguistics and Clinical Psychology, ACL Anthology.
[48] Benton, A., Mitchell, M., Hovy, D.: Multimodal mental health analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2017). https://doi.org/10.18653/v1/D17-1185
[49] Ji, S., Zhang, B., Wang, T., Wei, S., Yu, P.S. (2022). Mental health detection via social media: A survey. ACM TIST, 13(4), Article 69. https://doi.org/10.1145/3512730
[50] Ji, S., et al. (2021). Suicidal ideation detection via contextual embedding and hierarchical attention. Information Processing & Management, 58(3), 102542. https://doi.org/10.1016/j.ipm.2020.102542
[51] ibeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016). https://doi.org/10.1145/2939672.2939778
[52] Kaur, H., Mangat, V.: Depression detection on social media using machine learning and LIME. Journal of Ambient Intelligence and Humanized Computing (2020). https://doi.org/10.1007/s12652-020-01845-w
[53] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS) (2017). https://doi.org/10.48550/arXiv.1705.07874
[54] Yang, Z., Dai, Z., Yang, Y., et al.: Attention is all you need in NLP: Hierarchical attention networks for document classification. NAACL (2016). https://doi.org/10.18653/v1/N16-1174
[55] Jain, S., Wallace, B.C.: Attention is not explanation. NAACL-HLT (2019). https://doi.org/10.48550/arXiv.1902.10186
[56] Aldarwish, M., & Ahmad, H.F. (2017). Predicting depression levels using social media posts. IEEE ICT. https://doi.org/10.1109/ICT.2017.7976186
[57] Pestian, J.P., et al. (2017). Machine learning classification of suicidal and non-suicidal patients. Biomedical Informatics Insights, 10. https://doi.org/10.1177/1178222617725071
[58] Matero, M., et al. (2019). Suicide risk assessment with multi-level attention models. CLPsych Workshop. https://doi.org/10.18653/v1/W19-3009
[59] Sharma, E., & De Choudhury, M. (2021). Measuring and mitigating language biases in mental health classification. CHI. https://doi.org/10.1145/3411764.3445423
[60] Bentum, J., et al. (2022). Combining attention and SHAP for interpretable mental health detection. Expert Systems with Applications, 198, 116792. https://doi.org/10.1016/j.eswa.2022.116792
[61] Liang, H., et al. (2023). Hybrid explainability methods for transformer-based mental health detection. Information Sciences, 626, 441–456. https://doi.org/10.1016/j.ins.2023.03.016
[62] Anonymous ReDSM5 Study (2025). Clinically annotated DSM-5 symptom detection with explainable models. ArXiv Preprint (under review)

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)