International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 10 (2026) 93–111

Paper Type: Original Article

A Survey of Mental Health Intent Recognition Approaches

Mennatullah Eldakhakhni [1,*], Marwa Abdellah [1] and Mohamed Fouad [2]

[1] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig 44511, Egypt; Emails: m.mamdouh021@fci.zu.edu.eg; MMAboelazm@fci.zu.edu.eg.

[2] Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt; mohamed_mostafa@aast.edu.

Received: 12 Dec 2025 **Revised:** 04 Jan 2026 **Accepted:** 06 Feb 2026 **Published:** 07 Feb 2026

Abstract

As mental health conditions are increasing around the world, digital interventions are desperately required in a scale-form. The main focus of these solutions is intent recognition, the capacity to understand user intentions, whether it is a particular support request or indications of suicidal thoughts. The paper will overview the technical history of the area, beginning with the early machine learning methods and concluding with the current state-of-the-art transformer models, discussing the important datasets, multimodal methods and data ethical issues. One of the major issues of our analysis is the so-called deployment gap. We point to a significant mismatch: where NLP models are commonly found to be over 90 percent accurate in carefully controlled settings, they are often found to fail to do so in the real clinical scenarios. Exploring this discrepancy, we claim that the practical utility of academic metrics ought to be prioritized to responsible AI. It is a singularly thorough work that provides a six-decade historical point of view (between ELIZA and LLMs), a comparative analysis of methodologies and actual performance data, and a realistic assessment of the constraints that prevent its use in clinical settings.

Keywords: Mental Health, Natural Language Processing, Intent Recognition, Deep Learning, Transformers, Ethical Considerations, Deployment Challenges.

1 | Introduction

1.1 | Background and Motivation

It has been found that mental health disorders are now affecting more than a billion individuals in the world and depression and anxiety have become the leading causes of disability in the world. Since conventional clinical infrastructure, by definition, is incapable of scaling to such a demand on a population level, digital solutions that utilize natural language processing (NLP) have become crucial alternatives [3] to this type of demand. In this scenery, intent recognition has a distinctive clinical significance. As a process, it is much more complicated than regular text classification and involves the identification of small, sometimes latent hints such as suicidal thoughts and emergency situations and the emotional discomfort and ambivalent help-seeking patterns. The technical issue is further complicated by the fact that the discourse of mental health is semantically stratified, emotionally connoted, and extremely influenced by the demographic and cultural environment. The technical challenge is considerable: mental health language is semantically complicated, emotionally colored, and differs radically, according to demographics, cultures, and clinical situations [4].

Corresponding Author: m.mamdouh021@fci.zu.edu.eg

The reason behind this survey is a critical observation of the past. The discipline has gone through three various eras of technology:

- the Traditional ML Era (1960s2010s) characterized by rule-based and statistical methods.
- the Deep Learning Era (2013–2018) that proposed automated feature discovery.
- Transformer Era (2018): Transfer learning Pre-trained language models (2018) [1].

However, there is an underlying paradox, that even though one of the waves has better performance than the rest, none of them have been able to fully close the gap between the research measurements and effective clinical practice.

1.2 Research Questions and Survey Scope

To fill this gap, this survey will deal with four combined research questions:

- RQ1: How do the most significant changes in the methodology of mental health intent recognition between the conventional ML and recent transformer architectures change? What are the performance implications of each?
- RQ2: How do various approaches compare with each other in terms of strength, weaknesses and performance trade-offs and how do these apply to the actual clinical practice?
- RQ3: What are the key issues and constraints associated with datasets, evaluation and real-world implementation and what are the reported failures?
- RQ4: What are the most urgent ethical issues and future research directions of creating responsible, helpful mental health technologies with cultural-adaptation and equitable-access focus?

Scope and Differentiation:Although the given work mostly deals with English-language text models, it does not deny the need to conduct multilingual and multimodal research which is of essential importance. We also give certain emphasis on the gap between research and practice, emphasizing on the issues of measurement and real-world failures that most of the literature ignores.

Unlike recent surveys where the review is very limited, i.e. to transformers (e.g., Greco et al.),[1] this survey provides an extensive:

1. Historical background of 6 years.
2. Critical failure analysis.
3. Clear execution of performance in approaches.
4. Decent reporting on deployment issues.

2 | Historical Context: From ELIZA to Transformers

2.1 | The Long History of Mental Health Computing

The history of computational mental health analysis is much older than it is commonly thought.

Rule-Based Era(ELIZA and CDSS): The discipline commenced successfully in the 1960s with Weizenbaum's ELIZA[2], which employed simple pattern matching to imitate a Rogerian psychotherapist. Nonetheless, even at its basic level, ELIZA had a remarkably insightful result, which remains a noticeable impact in AI ethics: users are remarkably eager to project emotions onto the work of a machine[2]. Subsequently in the 1980s and 90s work was done on Clinical Decision Support Systems (CDSS)[5] that attempted to encode psychiatric knowledge in rule based logic. Although these systems provided a good interpretability, they were fragile enough to absorb the rich and varied differences and subtleties of real patient language [5].

Traditional ML Era (2000s2010s): With supervised learning becoming increasingly popular, the emphasis was placed on feature engineering by hand. To train classifiers, like SVMs[7], researchers painstakingly devised features, including bag-of-words, n-grams, emotion lexicons [6] and so on. Whereas the accuracy of these methods was respectable during benchmark performance (6580%), they eventually reached a semantic ceiling. They usually missed implicit cries of help, sarcasm and indirect negativity by using superficial indicators (such as negative keywords). Besides, the use of human intuition to develop features complicated the process of scaling.

2.2 | Deep Learning Revolution (2012–2018)

Deep learning was a breakthrough that left manual feature engineering far behind as the neural networks found hierarchical patterns themselves in raw data.

Key Architectures:

1. Sequential Modeling (RNNs): Architectures such as LSTMs and GRUs became the framework of choice to model temporal dynamics, e.g. the mood progression within a conversation[8]. LSTMs more particularly resolved the vanishing gradient issue, enabling models to acquire long-term dependencies, and attain F1-scores of approximately 82% in multimodal depression tasks[9] .
2. Feature Extraction (CNNs): CNNs are also computer vision inspired; they were good at identifying particular keywords of crises or linguistic indicators in the local windows of text[10]. They typically were however unable to model the larger sequential context in the same manner as RNNs.
3. Hybrid Models: By combining the merits of CNNs (local keyword detection) and RNNs (sequential context), researchers tended to have 1015 percent higher performance than single-architecture models.
4. The Trade-off: Although deep learning advanced performances to the 7585 percent range, it also created a major clinical issue the Black Box problem. As compared to previous models, these neural networks were much less interpretable, which made them difficult to apply in an environment where reasoning is required.

2.3 | Transformer Era (2018–Present)

The emerging of Transformers completely changed the scene. These models may be able to weigh the significance of every word in a sequence concurrently by substituting sequential processing with self-attention mechanisms, which will make complex dependencies more easily than ever before.

- Pre-training & Transfer Learning: It became a paradigm of first pre-training on large corpora (such as Wikipedia) and then fine-tuning on tasks. This method saved 80-90 percent of the requirement of the marked mental health data [3].
- State-of-the-art Performance: BERT and RoBERTa (performance) or DistilBERT (speed) models set the state of the art benchmark F1-scores to 85-95% [2, 12]. This was further narrowed down to MentalBERT, which uses psychiatric literature when training.

Key Models:

BERT (2018): Bidirectional encoder representations with 8090% F1-scores on mental health problems [1].

- RoBERTa (2019): The enhanced version of BERT that delivers the state-of-the-art performance on mental health classification [12].
- DistilBERT: Smaller and faster variant that can be used in real-time (100ms latency vs. 500ms full BERT)[13].

Domain-specific versions: MentalBERT, AraBERT are subsequently pre-trained on corpora in mental health[14] .

- The Reproduction of the Gap: Although these architectural revolutions were made, there is a paradox of critical nature. In practice, an average of 1525% reduction in performance has been found when going to practice, indicating that high accuracy scores in the laboratory do not imply clinical readiness.

3 | Related Work and Literature Overview

Table 1 presents comparative analysis of key studies across methodological waves, revealing a trajectory toward more complex, context-aware models while highlighting persistent challenges.

Table 1. Comparison of Mental Health Intent Recognition Approaches.

Study	Year	Focus / Methodology	Key Contributions	Findings	Limitations	Implications
Su et al. [16]	**2020**	**Scoping Review**	**Reviews of DL in mental health**	**Identifies trends; establishes DL relevance**	**Broad scope; not intent-focused**	**Motivates specialized surveys**
Le Glaz et al. [3]	**2021**	**Systematic Review**	**ML/NLP in mental health overview**	**Documents field evolution**	**Descriptive; lacks comparison**	**Shows methodological progress**
Shen et al. [9]	**2022**	**GRU/BiLSTM**	**Multimodal depression detection**	**F1=0.82; validates audio-text fusion**	**Small dataset; limited generalizability**	**Demonstrates multimodal benefits**
Mezzi et al. [17]	**2022**	**AraBERT**	**Arabic mental health intent recognition**	**+10–15% improvement over BERT**	**Language-specific; small dataset**	**Extends NLP beyond English**
Greco et al. [1]	**2023**	**Survey**	**Transformer-based models**	**Catalogs state-of-the-art**	**Narrow scope; limited pre-transformer**	**Establishes transformer dominance**
Khoo et al. [18]	2024	Systematic Review	Multimodal ML for mental health	Analyzes fusion benefits	Focuses on detection, not intent	**Establishes multimodal efficacy**

Main Observation of Table 1: An overview of these studies shows a very shocking pattern: whereas technical measures have significantly improved, the 6075 percentage band of classic ML is now 8595 percentage of new Transformers, there are no direct proportional increases in clinical success. This gap indicates that benchmark scores with norms could be a poor depiction of actual utility in the real world.

3.1 | The Research-Practice Gap: A Critical Observation

A literature review synthesis reveals a basic, frequently underestimated difference between research prowess and clinical efficacy.

- The Detection Accuracy Drop: Although transformer-based models show excellent accuracy in controlled benchmark settings, their operation often becomes worse once applied. The distributional changes and natural heterogeneity of realistic clinical populations that are significantly different to curated training data are largely responsible for this drop[4].
- Safety Failures in High-Stakes Scenarios: The most concerning issue is that conversational AI systems are disconnected. Although these systems perform well in theory, reported case studies have shown that these systems fail to identify explicit suicidal intent or escalation steps required when dealing with live interactions [4]. This failure points to the observation that high F1-scores on standardized datasets do not imply safety when used in the critical care setting.

- The Sensitivity-Specificity Dilemma: Deployment in the Real World enforces a dangerous trade-off. This is because optimization towards sensitivity (in order to identify all at-risk people) can quickly produce a tidal wave of false positives, causing clinician fatigue due to alarm and patient distress due to the false alarm[3]. Specificity on the other hand minimises false alarms but maximises risk of false negative- possibly missing a life threatening crisis.. It is this divide between the promise of research and the reality of practice that characterizes the existing situation in the field and delineates the fundamental challenges that the future work will have to tackle.

4 | Traditional Machine Learning Approaches

4.1 | Feature Engineering Fundamentals

Before the deep learning paradigm, mental health NLP was characterized by a need to use manual feature engineering, along with classical classification algorithms. The conventional process required a strict pipeline: text processing (tokenization, stemming) and transformation of linguistic data into numerical data, which would be processed by some algorithm, such as SVMs or Naive Bayes.

The main features representations:

- Statistical Features: Bag-of-Words (BoW) and TF-IDF are purely based on counting words[7], [19], as counts of word frequencies in text. Although good at topic modelling, the methods necessarily lost word order and long run context. To counter this, N-grams (sequences of neighboring words) were used by researchers, and they had been shown to be more effective at inference of local context than single-word keys.
- Lexicon-Based Approaches: The application of psycholinguistic dictionaries became the staple of this period. Such tools as LIWC (Linguistic Inquiry and Word Count)[6] enabled the researcher to map text to more than 80 psychological categories. Manual curation of domain-specific lexicons was also done to find signs of depression, including absolute words (e.g. always, never), negation patterns, strong utilisation of first-person pronouns, a known predictor of depression severity.

4.2 | Performance and Applicability

In the past, conventional ML methods provided classification accuracies within the 65 -80% range[3]. Advantages: The reason why these models are relevant is because they are highly interpretable; clinicians can easily examine the decision rules (e.g., this word raised the alarm). Also, their efficiency allows them to be used in privacy-sensitive and resource-constrained scenarios (Edge Computing).

Limitations: Limitations: Although, conversely, there was a major bottleneck in the form of manual engineering. Such models frequently do not have the ability to model long-range context or represent the granular semantic context that mental health narratives can contain e.g. sarcasm or concealed displays of distress. Modern Relevance: Although they have been almost completely replaced in complex multi-intent tasks by methods based on transformers,[1] traditional ML still can be an effective option in certain, lower-resource, cases where domain knowledge can be used to inform successful feature selection.

4.3 | Key Limitation: The Semantic Ceiling

This period in the history of methodology can be characterized as the "Semantic Ceiling. Manual features can be too inflexible to perform well to represent how language in mental health is context-dependent. Contextual ambiguity Contextual ambiguity A phrase such as I am fine is semantically ambiguous; it may reflect actual stability or hidden distress, based on the previous conversation history.

Symptom Overlap - Statement such as I am just tired may be a statement of symptom or somatic symptom of depression. This subtlety of context is often eluded by traditional classifiers that are limited to matching

keywords or local statistics. This failure to differentiate between different words with hugely different meanings is the major force behind the direction taken by the field towards the learned representations.

5 | Deep Learning Approaches

5.1 | Recurrent Neural Networks (RNNs)

In contrast to the traditional classifiers that processes words individually, Recurrent Neural Networks (RNNs) provided the internal memory concept so that the model is able to learn to capture time-varying dependencies through the state maintained as it processes words in a sequence.

- Addressing the Vanishing Gradient: Standard RNNs, in their turn, were unable to learn long sequences because of the so-called vanishing gradient problem, according to which learning signals decrease exponentially during backpropagation. This was a major drawback of mental health analysis, in which it is imperative to know the history of a patient or long-term mood variability as sequences are submitted word-by-word.RNNs do not lose this state (memory). The dependency among time steps can be modeled with the help of this memory[8].
- Modern Variants (LSTMs and GRUs): To overcome this, Long Short-Term Memory (LSTM) that incorporated gating (input, forget and output gates) to control information flow. This architecture was very effective in capturing long-range patterns, e.g. week-week trajectories of depression. Equally, Gated Recurrent Units (GRUs) provided a more computationally efficient form that had similar performance, and were appealing to real-time use[20].
- Bidirectional Context: More improvements resulted in Bidirectional RNNs (BiLSTMs), which handle both forward and backward sequence processing at the same time. This multimodal context modeling enabled BiLSTMs to reach F1-scores of approximately 82 percent on multimodal depression detection with a significant improvement over conventional ML[9].

5.2 | Convolutional Neural Networks (CNNs)

First developed to process images, CNNs were modified to NLP by considering text as a one dimensional grid.

- Mechanism: CNNs find local n-gram patterns by sliding filters (convolutions) over text windows and summarizing the result using pooling operations[11]. Clinical Utility: CNNs can be used effectively in the mental health setting, specifically in detecting phrases at the phrase-level e.g. detection of a certain crisis keywords, or detected phrases deemed to indicate immediate distress. Nevertheless, they tend to be worse at RNNs on tasks that need deep sequential learning or modeling a long-range context. Performance CNN-based methods perform similarly on specific tasks to RNNs, typically performing worse in sequential intent classification but better in phrase-level crisis detection.

5.3 | Hybrid Architectures: CNN-RNN Combinations

Rationale CNN-RNN Combinations CNN-RNN Hybrid architectures are based on the understanding that there are mental health signals at both the local (keywords) and the global (context) levels, so researchers introduced hybrids of CNNs and RNNs to exploit the strengths of these models. Architecture & Performance: These types of models use a CNN layer to extract salient local phrases which are inputted into an RNN layer to model the temporal relationships between them. The synergy generally produces 5-15 percentage point over and above single-architecture performance, allowing the system to recognize crisis keywords and appropriately understand them in the broader context of the conversation[11].

5.4 | The Deep Learning Breakthrough

The deep learning epoch reached its monumental breakthrough by removing the feature engineering crunch at a single stage and raising the performance to 75-85% (compared to 60-75% with conventional ML)[3]. The major benefits were learning hierarchical features and gaining better semantic processing. Constant Limitations: This increase in performance was however gained at a cost. Deep learning models also gave rise to the so-called Black Box problem: the opaque nature of the decision made by the model of the neural network, unlike the explicit decision rule of traditional classifiers, encodes no clear meaning to clinicians in high-stakes scenarios[16].

6 | Transformer-Based Architectures

6.1 | Self-Attention and the Transformer

The essence of the introduction of the Transformer architecture was to break the sequential processing (RNNs) by adding self-attention mechanisms[1]. This innovation enables the model to calculate importance weights of all positions of inputs at the same time.

- Computational Advantage: Transformers are able to massively parallel entire inputs, which greatly decreases training time. More to the point, they address the vanishing gradient issue that RNNs have, effectively learning long-range correlations despite the length of a sequence. Mental Health Context: Think about the following sentence: I never feel happy anymore. The relation between the negation (never) and emotion (happy) may be lost in a sequential model across a long distance[1]. On the contrary, in self-attention, the connection between these two words is explicitly brought out, and the model will identify the depressive mood and not misunderstand the word happy alone.

6.2 | Pre-training and Transfer Learning

Transformers made a two-stage paradigm popular which became the norm of modern NLP: 1. Phase 1 (Pre-training): Billions of parameters are trained on huge, unlabeled corpora (e.g., Wikipedia) to learn general linguistic structures. 2. Phase 2 (Fine-tuning): These pre-trained knowledge bases are then customized to particular tasks (e.g. detection of suicidal ideation) with smaller labeled datasets.

- Significant Advantage: This Transfer Learning technique saves 8090 percent of the labeled mental health data required to train by 8090 percent of that needed to train using raw data.

6.3 | Popular Pre-trained Language Models (PLMs)

Standard Bearers (BERT) BERT proposed bidirectional context processing. Its Base model (~110M parameters) can generally obtain 8090% F1-scores on mental health classification tasks, and fine-tuning on standard GPUs can take as short as 3060 minutes[21].

- Optimized Performance (RoBERTa): RoBERTa (Large, -355M parameters) regularly optimizes the training process of BERT, resulting in the state-of-the-art performance. Nonetheless, it has a better latency (approximately 1000ms/sample) and is therefore slower in real time application[12]. Efficiency-Focused (DistilBERT): where real-time performance is required, such as chatbots, the trade-off that DistilBERT provides is interesting. It can still perform 97 percent of the tasks of BERT with 40 percent less inference latency (approximately 100ms per sample), and is suitable in systems that require immediate response [13].
- Domain-Specific Variants: Domain and language-specific models such as MentalBERT (trained on psychiatric literature) and AraBERT (trained on Arabic) mitigate the issues of domain and language gaps with an improvement of 1015% over general-purpose models[14, 17].

6.4 | Computational Costs and Scalability Challenges

High accuracy is at a high cost of computation which poses a hindrance to its implementation.

- Resource Intensity: The models of high-performance (e.g. RoBERTa-large) are resource-intensive (you need at least 40GB of VRAM) and slow inference (500-100ms) which can be much higher than the tolerable latency of a conversation. Implication on deployment: This resource endowment poses accessibility obstacles on organizations operating in low-income areas. Moreover, because of the offloading of the computational workload to the cloud, it brings about privacy risks in sensitive patient data. A trilemma between speed, accuracy and privacy is therefore presented to organizations.

6.5 | Interpretability Through Attention Visualization

Self-attention weights can be visualized (e.g., heatmaps) to apparently indicate what words the model paid attention to when making its prediction e.g. harming or persistent. The Critical Limitation: Via intuition, research indicates an illusion of interpretability. Attention weights are not necessarily a reflection of human reasoning; a model can especially focus a lot on words that appear neutral and rely on the complicated, non-linear interactions of the features unavailable to the user. Attention maps should not be used as a final answer to the questions in clinical scenarios with high stakes [21].

7 | Advanced and Hybrid Approaches

7.1 | Hybrid Deep Learning Models

- Rationale: Hybrid models are created to combine the strengths of different architectures which are the usage of CNNs in the extraction of local features, RNNs in a sequencer process, and Transformers in long-range dependencies.

- Shared Architectures: A typical hybrid design It takes a CNN layer to detect salient linguistic cues and key phrases, which are then passed to a BiLSTM to capture the contextual dynamics of the story. Empirically, these combinations prove to have a 5-15 percent performance gain compared to single-architecture baselines [11]. Recent Developments: There has been a shift toward applying contextual embeddings (such as BERT) to specialized downstream architectures. Also, another important innovation [16] has appeared as Multi-Task Learning (MTL) in which one model is trained to perform similarly-related tasks (e.g. sentiment and severity and intent detection) enhancing the overall performance via shared representation learning.

7.2 | Multimodal Approaches

Mental health occurs via numerous communicative mediums, and the analysis of the text is not sufficient to address this issue. Multimodal machine learning is a solution to this issue because it involves the combination of various sources of data to provide a more holistic evaluation [18].

- Linguistic Modality (Text): This is the most studied source of data, as it is based on posts on social media, chat histories, and clinical records.

- Acoustic Modality (Audio): Vocal patterns are important biomarkers that are usually non-visual when using text. Key features include: o Pitch: Pitch is often linked with depression and a decrease in variability. o Speech Rate: the depression has psychomotor retardation that normally leads to slow speech. o Voice Quality: Hoarseness, tremor, and flat prosody.

- Performance Impact: It has been shown that although the audio analysis alone has 60-70% accuracy in detecting depression, the combination with textual data improves the accuracy to the 80-85% range[9] regardless of the cross-modal integration, which is effective.

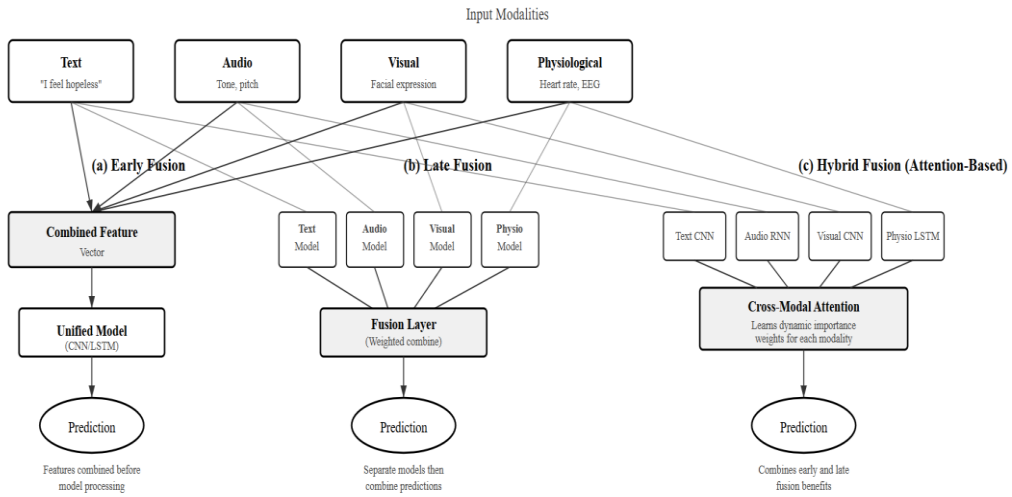

Figure 1. Multimodal Fusion Architectures for Mental Health Assessment.
Source: Developed by the Authors

- Visual Modality (Video): This analyzes non-verbal communication like facial expressions, eye contact, and movements of the head. Although the use of multimodal data with video has shown accuracy of depression severity classification of about 79 percent [22], the consideration of visual data creates serious privacy implications that make it difficult to deploy.
- Physiological Modality (Sensors): Wearable technology allows monitoring the biomarkers of heart rate variability, skin conductance and sleep patterns. These signals provide a longitudinal view which exposes temporal patterns of misery that cannot be seen in just one interaction [18].

7.3 | Multimodal Challenges and Trade-offs

In spite of their potential, multimodal approaches have serious impediments to deployment:

- Data Alignment: It is a computationally hard problem to synchronize asynchronous streams of data (e.g. match a textual phrase with a brief micro-expression). Missing Modality Robustness: Real-life systems should be able to deal with partial data (e.g., a user switching off his or her camera). The models should be constructed in such a way that they do not fail but promote graceful degradation to single-modality. The computational load Multimodal processing generally adds 2-5x inference latency, which is difficult to measure in real time.
- Annotation Bottlenecks: Multimodal data collections are costly to produce, requiring interdisciplinary skills (clinical, linguistic, and signal processing) and leads to a lack of data[18].

7.4 | Domain-Specific and Culturally Adapted Models

The generic language models are not always able to reflect the specifics of the mental health discourse that differs considerably across cultures.

- Domain-Adaptive Pre-training (DAPT): To address the vocabulary gap, models such as MentalBERT are additionally pre-trained on psychiatric literature and forums and do better on domain-dependent tasks than general BERT models [14].
- The Arabic Gap: Dealing with linguistic underrepresentation, AraBERT (optimised to mental health) has shown a 1015% better performance than multilingual baselines in Arabic-speaking populations [17].

The Imperative of Cultural Adaptation: The studies have shown that literal translation does not make do. Efficient AI interventions should support Eight Critical Dimensions of Adaptation

such as:

(1) Language idioms.

(2) Gender/Authority roles.

(3) Cultural metaphors.

(4) Relevant content scenarios.

(5) Culturally specific ideas of distress.

(6) Valued outcomes/goals.

(7) Therapeutic methods.

(8) Socio-political context.

The Digital Mental Health Divide: There is a serious lack of imbalance regarding data availability. Whereas English datasets possess millions of examples, other languages such as Arabic, Hindi and African dialects contain hundreds at best or no examples whatsoever. The result of this discrepancy is the Digital Divide, in which non-English speakers are segregated to using proven inferior AI mental health products[1].

8 | Datasets, Training Paradigms, and Evaluation

8.1 | Key Mental Health Datasets

Mental health NLP development depends critically on dataset quality and availability as shown in Table 2:

Table 2. Key Datasets for Mental Health Intent Recognition.

Dataset	Source	Size	Annotation	Performance	Key Characteristics	Known Biases
DAIC-WOZ **[23]**	Clinical	189 sessions, 960 min	Depression severity (PHQ-8)	F1: 0.85–0.95	Multimodal (audio, video, text); high-quality	80% Caucasian, 51% female; demographic skew
CLPsych **[24]**	Social media	2,000–5,000/year	Depression, PTSD, severity	F1: 0.70–0.85	Annual shared tasks; variable focus	Predominantly English; platform-specific biases
C-SSRS **[24]**	Social media	5,000 posts	Suicide risk (0–5 scale)	Varies	Reddit data; crowdsourced annotation	High annotation ambiguity; inter-rater disagreement 0.45–0.65
Emotional Audio-Textual **[9]**	Clinical	500 sessions	Depression (binary)	F1: 0.82	Audio-text pairs; balanced dataset	Limited ethnic/cultural diversity
AVEC 2014 [25]	Clinical	Video interviews	Depression severity	Acc: 79%	Multimodal (video, audio); structured interviews	Limited demographic diversity
E-DAIC**[22]**	Clinical/Crowd	189+ sessions	Depression severity	F1: 0.86–0.89	Extended DAIC with crowdsourced annotations	Annotation noise; demographic skew
Uncurated Public Repositories **[4]**	Mixed	Highly variable	Inconsistent	Highly variable	Often poorly documented	Highly variable quality; many synthetic/preprocessed

Key Insight: The Quality-Quantity Trade-off A literature review develops a principle of mental health NLP:

- Goodness of data sets is a more influential indicator of model effectiveness compared to amount of data. In practice, a clean dataset of 500 clinical sessions can usually be more generalized compared to a 50,000-post social media dataset that is noisy. This dominance has four essential dimensions, namely annotation rigor (expert vs. crowd), label accord, demographic representativeness, and domain specificity.

8.2 | Critical Issues in Mental Health Datasets

- Scarcity vs. Noise: It is exceptionally scarce to find high-quality, clinically validated data. Although social media data is highly available, it poses a high level of noise and systemic bias. As an example, dataset hosted on Reddit is severely over-representative of younger, English-speaking people (average age is approximately 30) and thoughtfully marginalizes older people, non-native speakers, and other people in extreme crisis that cannot describe their affliction on the Internet[4]. Demographic Bias: Despite the clinical data being of gold standard, striking skew can be observed. Caucasians [23] occupy 80% of the DAIC-WOZ, which is popularly used to make benchmarks. Models that are developed using such homogeneous data will have a high risk of reproducing health disparities, especially with black patients, historically underdiagnosed[3] .
- Data Leakage: Data leakage is a widespread problem whereby, information in test sets accidentally leaks into training data. This is especially acute when it comes to time-series analysis because an inability to adhere to the rigid temporal limits artificialistically inflates performance metrics[4].
- Ethical Limitations: The application of social media information maneuvers around a very precarious ethical situation as relates to privacy and consent. When researchers compute the sensitive disclosure-related information, there should be a balance between the utility of the data and the necessity to achieve informed consent in the first place.

8.3 | Training Paradigms

Supervised Learning: This is the conventional method, which provides obvious learning indicators, but requires human annotation, which is laborious and costly. Transfer Learning: Has become the new paradigm of Transformer based models. This strategy can cut down on the need to use labeled mental health examples by 8090 percent by the pre-training phase of general corpora and fine-tuning on task-specific data [1] . Multi-Task Learning (MTL): Multi-task to learn one model with multiple goals (ex: predicting intent, sentiment, and severity) have demonstrated to generalize better (by 5-10 percent) than single-purpose models trained on the same objectives [19].

8.4 | Training Paradigm Limitations

Bias in Data Augmentation Data augmentation methods are susceptible to infuse new clinical semantics, such as those based on back-translation. As an illustration, when it comes to back-translation of I am thinking about hurting myself, the result could be I am thinking about self-harm. Although the meaning of words is close, the change of urgency and wording may have a dramatic diagnostic effect. Synthetic Data Validity Generative models (GANs, LLMs) may present a tempting solution to the problem of data scarcity, although it has been found that synthetic mental health data tends to reproduce the chaotic, non-linear nature of actual clinical presentations. Overfitting: Since, even the largest annotated mental health data sets use thousands of examples (compared to the millions of examples in general NLP), models are highly prone to overfitting, and learning the artifacts of the training set instead of generalizing to new patients.

8.5 | Evaluation Metrics for Mental Health

The Class Imbalance Issue Standard measures such as Accuracy are misleading when applied in mental health. As suicidal ideation is usually less than 5 percent of the data, a model that predicts No Risk of all people gives 95 per cent accuracy but is clinically meaningless. Assessment should hence focus on Precision, Recall, F1-Score and AUC. Cost-Sensitive Assessment: The misclassification mistakes are asymmetrical. True Negative (excluded a suicidal person) is possibly life-threatening, whereas True Positive (false alarm) leads to a distress but life-saving. The literature therefore recommends that the weight of False Negatives be 10-100 times greater than that of the False Positive in risk assessment model[15]. The Clinical Evaluation Gap: There is a major disjunction between research and practice. The majority of the studies measure models only on the basis of retrospective criteria (e.g., CLPsych Shared Tasks). There is almost no prospective clinical trials that is the gold standard of medicine. Models that are tested in real-world deployment are often found to have a 1525% reduction in performance over their published benchmark performance [3].

9 | Applications and Real-World Deployment

9.1 | Key Application Areas

Chatbots and Virtual Agents: Chatbots and Virtual Agents are immediate, scalable, and anonymous assistance, and in many cases, act as a front line defense mechanism. Their range of functionality was emphasized in a survey by Abd-Alrazaq et al., yet empirical validation is paramount. In particular, a randomized controlled trial (RCT) by Fitzpatrick et al. (2017) showed that the use of the automated agent, such as Woebot, delivering Cognitive Behavioral Therapy (CBT), led to a much more significant decrease in symptoms in young adults as opposed to a control group that used standard self-help materials [26].

Social Media Surveillance: In addition to direct contact, systems are also implemented in the form of passive identification of risk indicators in social posts. Studies have indicated that it is possible to flag patients at risk of self-harm in collaborative care models, and thus reach out to them proactively [27]. Clinical Decision Support Systems (CDSS): The systems are designed to help clinicians monitor conditions by analyzing transcript and notes of interviews. Though practical in a controlled setting, their validity tends to falter in a real-life situation since there is a considerable variance in how various clinicians record their interactions with patients [5].

9.2 | Documented Deployment Challenges

Irrespective of the accuracy that is high on research, when these models are transferred to production it has been observed that they will fail in critical ways: The "Reality Gap" (Performance Degradation): It is a well-known fact that models, which perform to state-of-the-art on benchmark, often experience a 15-25% decline in performance on deployed clinical workflows [3]. Risk Escalation Dilemmas: Early detection systems do not have a way of calibrating the proper intervention protocols. They are prone to make mistakes at both ends: being too conservative (and therefore failing to intervene in time) or being too pushy (resulting in over-causation of alerts that lead to clinician alarm fatigue).

- Demographic Inequality: There is a high standard deviation of performance between different populations in models. The evidence records much lesser accuracy among the demographic groups that are underrepresented, and this raises the question of equitable care [4]. Using a distribution shift, generalization failures can occur: When applied to an alternative hospital system, or language, or even care setting, a model trained on a particular clinical population tends to fail.

9.3 | Interpretability and Clinical Trust Issues

- BlackBox Barrier: To trust an AI assessment, a clinician must not only be accurate but must also get an explanation. Deep learning model opacities make it difficult to adopt, as clinicians require information regarding why a patient was raised as high-risk, i.e. what phrases or behavioral patterns were used to raise the red flag [1].
- Possible Attention Mechanism limits: Although Transformer attention weights can be visualized as a measure of explainability, this is not always a clinically sound approach. What the model may do is to score high on attention of seemingly neutral words but use complex, non-linear interactions between features that are not visible to the human viewer as the final determinant of the actual decision the model takes. This poses a threat where the visualizations give a misleading perception of the knowledge [1]. 10 | Ethical Reflections and Gaps that are Crucial.

10 | Ethical Considerations and Critical Gaps

10.1 | Privacy and Data Security

- Mitigation Strategies: Mental health information is sensitive, and that is why it is important to adhere to the requirements of such regulatory frameworks as GDPR and HIPAA. Critical technical protective measures are strict data anonymization/de-identification, detailed informed consent procedures, and strong encryption procedures.
- Utility -Privacy Tension: One of the most commonly debated issues seems to be the necessity to balance the data utility and the maintenance of privacy. Despite highly developed anonymization, the individual storytelling that is inherent to a disclosure of mental health poses continuous threats of re-identification, making it complicated to share datasets to research.

10.2 | Bias and Fairness

- Risk of Amplification: AI models are extremely prone to learning and enhancing the systematic bias in the training data, which is likely to result in potentially harmful health disparities. Demographic Bias: The standard dataset called DAIC-WOZ consists of 80 percent Caucasians and 51 percent women [23].
- Models that are trained using such biased data will have poor performance among non-Caucasian, male, and non-binary groups.
- Language and Age Bias: Models which are only trained on the English data do not work at all with non-native speakers. In the same way, the models that are trained on social media websites such as Reddit (average age is around 30) [4] would be inaccurate in the symptoms of geriatric depression which appear in different ways.
- Underdiagnosis Disparity: Studies have directly recorded the problem of increased underdiagnosis in Black individuals where systems with Caucasian data, which are predominantly used in health care systems, are involved, continuing to perpetuate pre-established healthcare disparities.
- Pathways to Mitigation: To mitigate this, it is necessary to use a multi-faceted strategy, involving taking an active part in all the audience datasets auditing, employing fairness-conscious machine learning methods, and a collective effort to gather a variety of diverse and representative data.

10.3 | Consequences of Misclassification

- In a mental health setting, classification errors have asymmetric and dire effects which cannot be measured through standard measures (such as Accuracy).

- False Negatives (The Fatal Error): The loss of a potentially dangerous person (e.g., the inability to detect suicidal thoughts) may lead to a fatal outcome. According to cost-benefit studies, when conducting suicide risk assessment, the weight of true negative is supposed to be 50-100x that of the true positive[15]. False Positives (The Distress Error): These are false warnings being sent to a user that they are high-risk that trigger unnecessary distress, possible stigma, and unwanted interventions (e.g., involuntary hospitalization), which may undermine user trust and deter further help-seeking.
- Ethical Stake: Although a false positive can cause some temporary pain, a false negative can be a permanent one.

10.4 | Interpretability-Criticality Relationship

Important Result: An inverse correlation frequently exists between interpretability of a model and the clinical risks of its use.

- Low-Stakes: In general mood assessment, it might be acceptable to use black-box transformers.
- High Stakes: Interpretability is needed in the case of suicide risk determination. Nevertheless, such high-stakes problems tend to use the least interpretable deep learning models. Implication Clinical deployment choices should be restricted to interpretability requirements. A explainable model with lower accuracy is ethically desirable in a crisis than an opaque model with greater accuracy in other situations.

10.5 | Human-in-the-Loop as Ethical Imperative

- Principles: AI Systems should not be developed to substitute human clinicians but serve to complement them. Design Constraints: Human intervention has to be activated at all time when making critical decisions which includes: identifying a suicide risk or crisis escalation. Psychiatric holds or Forced Hospitalization Systems should not be given the freedom to make psychiatric holds or forced hospitalizations.
- Safety Procedures: The users should not lose the power to escalate to the human clinicians. Moreover, when the system uncertainty is high, one of the consequences should be automatic handover to the human being instead of automatic reaction.
- Implementation Issues: There is a significant conflict between the need to have scalable, automated systems and the budgetary and burden requirements of having human-level control [3].

11 | Synthesized Insights: Meta-Analysis of the Field

Key Finding: Model performance when using research standards is found to decline uniformly and empirically when deployed in the real world.

- The “Reality Gap: A model with F1-score of 0.90 on a fixed benchmark will tend to decrease to the 0.65-0.75 range in clinical practice- a 15-25 percent degradation [3].
- Root Causes: 1. Distribution Shift: Research material is usually based on a convenience sample (e.g., active users of the Reddit platform) not representative of the much larger, more silent clinical population [4]. 2. Measurement Artifacts: The clinical notes used in the real world are in marked contrast to those of research protocols, which are standardized. 3. Temporal Leakage: Benchmarks resort to randomized train-test splits, which reveal future information accidentally. True prospective prediction is much more difficult and is required to be deployed. Implication: Benchmark scores as published must be considered to reflect upper-bound theoretical limits as opposed to deployment preparedness. The realistic expectations should be brought down by approximately 20 percent.

11.1 | The Performance-Generalization Trade-off

Misconception: One of the most commonly held beliefs is that newer architectures are always better (i.e. Transformers are always better than RNNs, which are always better than Traditional ML).

- The Subtlety of the Reality: The choice of methods should be determined by task parameters and not the novelty of time.
- Task Dependency: In the case of tasks such as crisis keyword spotting, traditional ML using curated lexicons may be faster, interpretable and equally accurate than deep learning.
- Data Dependency: o Small Data (Less than 1,000 examples): Deep learning does not tend to do well compared to traditional ML because of reduced overfitting risks [16]. o Medium Data (~10,000 examples): It is the dominance of transformers. o Big Data (more than 100,000 examples): It is apparent that transformers are better. Implication: A researcher is to avoid going to default state-of-the-art Transformers without evaluating the demands of data volume and interpretability.

11.2 | Data Quality Over Quantity

Critical Result: Mental health NLP follows a different pattern where quality rather than quantity is the primary determinant of successful results, as opposed to general NLP in which more data is better.

- Evidence: DAIC-WOZ dataset (189 sessions only, but clinically validated) reliably provides models with higher F1-scores (0.85 -0.95) [23] than models trained on the CLPsych dataset (thousands of noisy Reddit posts, F1 0.70 -0.85) [27]. Implication: The discipline needs to shift towards scraping large, noisy social media volumes to investing in smaller, annotated and clinically validated corpora.

11.3 | The Cultural-Linguistic Divide

Important Result: NLP Non-English mental health research is currently approximately 3-5 years behind English-centric studies in terms of methodological sophistication [17].

- The Data Disparity:
    - English: Millions of training examples are available.
    - Spanish: Thousands.
    - Arabic/Chinese/Hindi: Hundreds.
    - African Languages: Nonexistent practically.
- Consequence: This causes a Digital Mental Health Divide, in which speakers of non-English languages are put into use of demonstrated inferior AI tools that are culturally insensitive. 3. New implications: Simple translation is not enough, subsequent funding and research should focus on the development of native, culturally-specific datasets in languages underrepresented.

11.4 | The Interpretability-Criticality Relationship

Significant Result: There is an inverse correlation between the interpretability of the model and the clinical assets of the application.

- The Paradox: Those tasks that are the most explainable (such as suicide risk assessment) require the most explainable models to be used but are usually addressed with the least explainable ones (deep neural networks) [1]. Implication: Interpretability should be used to limit clinical deployment decisions. In high-risk decisions, an open model that is a little less accurate is safer and more ethical to a highly-accurate black box that is inaccessible to clinicians.

12 | Future Research Directions

12.1 | High Priority (Next 2–3 Years)

- To Clinical Trials: The discipline needs to immediately shift away and move past retrospective benchmark assessments, and move on to Randomized Controlled Trials (RCTs) [3]. The creation of safety and effectiveness in live settings is uncompromising. The funding bodies and journals must therefore focus on deployment-oriented research as opposed to papers that only suggest gradual architectural modifications.
- Between Correlation and Causality: The existing models are essentially reactive, in that they signal symptoms once they begin. The Causal Modeling would have to be incorporated in future systems to determine the causal factors behind mental health conditions. Things that answer why and how of progression are a mandatory prerequisite to the transition of passive surveillance into proactive action.
- Stability to Distribution Shifts: The second significant technological issue is that it must be stable to new populations, or languages, when models are presented with novel populations. Studies should be conducted on theoretical developments in Domain Adaptation to ensure that a model that has been trained on one set of environments does not crash in another.

12.2 | Medium Priority (3–5 Years)

- Privacy-Preserving Personalization: The following generation of models should address two competing objectives: to form a deep, stable knowledge about the individual user (Personalization) and preserve his or her identity strictly (Privacy). It would necessitate the incorporation of such methods as Differential Privacy into the training pipeline to avoid re-identification attacks [4].
- Safety-Constrained Generative Intelligence: Large Language Models (LLMs) are highly dangerous, although they have unparalleled empathetic abilities. The future work should consider how to utilize their conversational subtlety when it comes to enforcing what are called Hard Constraints mathematical guardrails to ensure injurious advice or poisonous output despite the timing.
- Globalization of Mental Health AI: The English-centered systems should be left behind. This includes the creation of Cross-Cultural Frameworks that are not just the translation of words but encode the cultural differences in the way mental distress is manifested and interpreted across the world[17]).

12.3 | Long Term (5+ Years)

- Refining vs. Reinventing: New architectures are of scientific interest, but today they are not as important as the validation of existing tools. The greatest clinical benefits over the next decade are probably not due to a new invention of a superior Transformer, but to the quality of data, enhanced clinical interconnectedness and achieving the perfection of the already existing state-of-the-art models.

13 | Conclusion

13.1 | Synthesis of Evolution and the Deployment Gap

- Following the line of mental health intent recognition progresses over the last six decades, one can observe an outstanding development. The era of rigidity, rule-based logic of the 1960s (ELIZA) and manual feature engineering of the 2000s have been supplanted by the current era of high-performance Transformers. Although this change has opened a kind of linguistic insight that was

previously believed unattainable, a major paradox has been pointed out in the course of our review: the Deployment Gap. Though accuracy measures in research papers are often above 90, such numbers do not always translate into clinical utility, and the numbers dramatically decrease when models are applied to a complex set of interactions with patients.

13.2 | Persistent Barriers

- Numerous reasons provide the explanation of this unconnection. To start with, the discipline has been characterized by use of retrospective standards instead of prospective clinical tests, which is the best in the medical field. Second, the models have a generalization weakness in that they usually fail when applied in alternate populations, languages, or health systems. Probably most importantly, the black box quality of the modern Deep Learning leads to a lack of trust; critical clinical decisions need to be explainable, and modern architectures are not well equipped to achieve it. Moreover, there is still a very high level of a so-called digital divide in which non-English populations are not adequately served by the existing NLP developments.

13.3 | The Path Forward

To fill this gap, a multi-pronged strategic change is necessary:

1. Focus on Prospective validation: Evaluation should be taken to serious clinical trials as means of measuring real-world results rather than static test-sets.
2. Mandate Interpretability: In high stakes decisions, the model selection must be restricted due to explainability considerations, at the cost of a small cost in accuracy.
3. Adopt Human in the Loop Design: Systems should be designed in a way to enhance clinician knowledge as a decision support system, but not to substitute it.
4. Invest in Data Quality & Diversity: Future actions should focus on curating diverse clinically annotated datasets, as opposed to accumulating volume of huge and noisy social media data.
5. Operationalize Cultural Adaptation: Understanding that translation is not enough; systems that work entail strong cultural adaptation that incorporates the local communities.
6. Impose Safety Constraints: Deployment must have hard constraints and guardrails which mathematically guard against outputs that are harmful. Finishing Vision: The end vision is not to create more intelligent algorithms, but to create an open ecosystem in which AI supports human empathy and knowledge. To accomplish this vision, a long-term collaboration between computer scientists, clinicians, and ethicists is required to provide digital mental health care that is more effective, as well as more equitable and accessible worldwide.

Author Contribution

Mennatullah Eldakhakhni thought the research idea, technologically developed the survey design, carried the overall literature review and analysis, conducted formal analysis of the results and drafted the original manuscript. Dr. Marwa Abdellah and Assoc. Prof. Mohamed Mostafa Fouad were extremely helpful in offering guidance and justification of the research methodology and findings, revising and editing the paper to a level of scholarly rigor and clarity, as well as in managing project management during the research process. All the authors have seen and accepted the published form of the manuscript.

Funding

This research received no funding.

Data Availability

The data sets that were used during this study were fully anonymized and publicly available. This is a review of literature and the methods that have been used; no new data had been produced. All the datasets used in the analysis of the study are mentioned in the references section of the manuscript and publicly available within their repositories.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Greco, C. M., Simeri, A., Tagarelli, A., & Zumpano, E. (2023). Transformer-based language models for mental health issues: A survey. Pattern Recognition Letters, 167, 204–211. https://doi.org/10.1016/j.patrec.2023.02.016

[2] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168

[3] Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... & Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. Journal of medical Internet research, 23(5), e15708. https://www.jmir.org/2021/5/e15708/

[4] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. npj Digital Medicine, 3(1), 43. https://doi.org/10.1038/s41746-020-0233-7

[5] Musen, M. A., Middleton, B., & Greenes, R. A. (2021). Clinical Decision-Support Systems. In E. H. Shortliffe & J. J. Cimino (Eds.), Biomedical Informatics: Computer Applications in Health Care and Biomedicine (pp. 795–840). Springer International Publishing. https://doi.org/10.1007/978-3-030-58721-5_24

[6] Pennebaker, J., Francis, M., & Booth, R. (1999). Linguistic inquiry and word count (LIWC).

[7] Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

[8] Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

[9] Shen, Y., Yang, H., & Lin, L. (2022, May). Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6247–6251). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746569

[10] Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018, December). Convolutional Neural Network (CNN) for Image Detection and Recognition. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 278–282). IEEE. https://doi.org/10.1109/ICSCCC.2018.8703316

[11] She, X., & Zhang, D. (2018, December). Text Classification Based on Hybrid CNN-LSTM Hybrid Model. In 2018 11th International Symposium on Computational Intelligence and Design (ISCID) (pp. 185–189). IEEE. https://doi.org/10.1109/ISCID.2018.10144

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692

[13] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. https://doi.org/10.48550/arXiv.1910.01108

[14] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022, June). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7184–7190). European Language Resources Association. https://aclanthology.org/2022.lrec-1.778/

[15] Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical metrics used to evaluate health care chatbots: scoping review. Journal of medical Internet research, 22(6), e18301. https://www.jmir.org/2020/6/e18301

[16] Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. Translational Psychiatry, 10(1), 116. https://doi.org/10.1038/s41398-020-0780-3

[17] El-Alami, F., Ouatik El Alaoui, S., & En Nahnahi, N. (2022). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences, 34*(10, Part A), 8422–8428. https://doi.org/10.1016/j.jksuci.2021.02.005

[18] Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors, 24(2), 348. https://doi.org/10.3390/s24020348

[19] Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1(1), 43–52. https://doi.org/10.1007/s13042-010-0001-0

[20] Dey, R., & Salem, F. M. (2017, August). Gate-variants of Gated Recurrent Unit (GRU) neural networks. In 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS) (pp. 1597–1600). IEEE. https://doi.org/10.1109/MWSCAS.2017.8053243

[21] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[22] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., ... & Pantic, M. (2017, October). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In Proceedings of the 7th annual workshop on audio/visual emotion challenge (pp. 3-9). https://doi.org/10.1145/3133944.3133953

[23] Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. (2014, May). The distress analysis interview corpus of human and computer interviews. In Lrec (Vol. 14, pp. 3123-3128).

[24] Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018, June). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 25–36). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0603

[25] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., & Pantic, M. (2014, November). AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (pp. 3–10). Association for Computing Machinery. https://doi.org/10.1145/2661806.2661807

[26] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 4(2), e7785. https://mental.jmir.org/2017/2/e19

[27] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015, June). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 31–39). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1204