www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92

e-ISSN : 2583-1062
Impact Factor : 7.001

AI-POWERED MENTAL HEALTH DETECTION SYSTEM USING TEXT ANALYSIS

Dipak Nitin Narkhede[1], Dr. Santosh Jagtap[2]

[1,2]Prof. Ramkrishna More College, Pradhikaran, Pune, India.

Email: dipaknarkhede42920@gmail.com

Email: st.jagtap@gmail.com

ABSTRACT

Mental health disorders impact over 970 million people globally, with depression and anxiety ranking among the leading causes of disability (WHO, 2023). Research shows that more than 70% of individuals experiencing mental distress do not receive timely intervention, underscoring the need for AI-driven early detection systems to address this gap (Global Burden of Disease Study, 2022). This study focuses on the development and evaluation of AI-powered mental health detection systems using text analysis to enhance early intervention efforts. By analyzing digital communication data, including social media posts, clinical interviews, and crisis conversations, this research examines the potential of AI in identifying depression, anxiety, and suicidal ideation effectively.

1. INTRODUCTION

1.1 Background of the Study

Mental health constitutes a complex and pervasive global challenge, affecting millions of lives worldwide and often leading to severe consequences when left unaddressed. Approximately 12.5% of individuals experience mental health issues at some point in their lives1. Despite the prevalence of mental health disorders, traditional diagnostic methods remain limited by accessibility barriers, stigma, and delayed intervention.

The proliferation of digital communication platforms has created vast repositories of personal data that hold immense potential for mental health analytics. Social media platforms, messaging applications, and online forums have become spaces where individuals express their thoughts, emotions, and experiences—often revealing indicators of their mental wellbeing. This digital footprint presents a unique opportunity for developing automated systems capable of identifying early signs of mental health conditions12.

Artificial intelligence (AI) technologies, particularly those leveraging natural language processing (NLP) and machine learning, have demonstrated promising capabilities in analyzing textual data to identify patterns associated with various mental health conditions. These AI-based approaches offer the potential for scalable, accessible, and early detection systems that could complement traditional mental healthcare services6.

1.2 Problem Statement

Despite advancements in mental healthcare, significant challenges persist in the early detection and intervention of mental health conditions. Current diagnostic processes often rely on individuals actively seeking help, which may be hindered by stigma, lack of awareness, limited healthcare access, or financial constraints. By the time many individuals receive professional attention, their conditions may have progressed significantly, complicating treatment and recovery2.

Traditional mental health assessments typically depend on self-reporting and clinical interviews, which may be influenced by recall bias, social desirability bias, and variations in communication abilities. These methods also require trained professionals, limiting their scalability in addressing global mental health needs6.

Furthermore, the increasing prevalence of digital communication has created both challenges and opportunities in mental healthcare. While people increasingly express mental health concerns online, these digital signals often go undetected or unaddressed without appropriate monitoring systems12.

This research addresses the critical need for developing, implementing, and evaluating AI-based systems capable of detecting signs of mental health conditions through text analysis, potentially enabling earlier intervention and support for individuals in need.

1.3 Research Objectives

This research aims to accomplish the following objectives:

1. To develop and implement an AI-based system for detecting mental health conditions through text analysis using machine learning and deep learning approaches
2. To evaluate and compare the performance of traditional machine learning (SVM) and transformer-based models (BERT) in mental health text classification

3. To assess the accuracy, precision, recall, and F1-score of the developed models in identifying various mental health conditions
4. To analyze the practical applications and limitations of text-based mental health detection systems
5. To identify ethical considerations and best practices for implementing AI-based mental health detection tools

1.4 Research Questions

This study seeks to answer the following research questions:

1. How accurately can AI-based systems detect signs of mental health conditions through text analysis?
2. What are the comparative advantages and limitations of Support Vector Machine (SVM) and BERT models in mental health text classification?
3. Which features or patterns in textual data are most indicative of specific mental health conditions?
4. What ethical considerations must be addressed when implementing AI-based mental health detection systems?
5. How can text-based mental health detection systems be effectively integrated into existing mental healthcare frameworks?

1.5 Scope of the Study

This study focuses specifically on the development and evaluation of text-based AI systems for mental health detection. The scope encompasses:

- Analysis of textual data from online platforms (primarily Reddit mental health discussions, DAIC-WOZ interview transcripts, and Crisis Text Line conversations)
- Implementation and comparison of two distinct approaches: traditional machine learning (SVM with TF-IDF vectorization) and transformer-based models (BERT)
- Detection of common mental health conditions including depression, anxiety, and suicidal ideation
- Evaluation of model performance using standard metrics (accuracy, precision, recall, F1-score)
- Ethical considerations and guidelines for responsible implementation

The study does not attempt to provide clinical diagnosis or replace professional mental healthcare services but rather explores the potential of AI as a complementary tool for early detection and intervention.

1.6 Significance of the Study

This research contributes to the field of mental healthcare in several significant ways:

1. **Early Intervention**: Early detection of mental health conditions can lead to more timely interventions, potentially improving treatment outcomes and reducing severity4.
2. **Scalability**: AI-based detection systems could help address the global shortage of mental health professionals by providing scalable screening tools that can reach larger populations26.
3. **Accessibility**: Text-based detection systems could help identify individuals who may not otherwise seek help due to stigma, lack of awareness, or limited access to healthcare services6.
4. **Integration of Technology**: This research demonstrates how modern AI technology can be effectively integrated into mental healthcare frameworks, potentially transforming approaches to mental health monitoring and intervention3.
5. **Evidence-Based Development**: By rigorously evaluating different models and approaches, this study contributes to the evidence base for developing effective mental health detection systems4 5.

2. LITERATURE REVIEW

2.1 Introduction to Literature Review

The field of text-based mental health detection using artificial intelligence has evolved rapidly in recent years, drawing from advancements in machine learning, natural language processing, and mental healthcare. This literature review examines the theoretical foundations, methodological approaches, and existing research related to AI applications in mental health detection. The review synthesizes findings from over 20 scholarly sources, identifying current knowledge, best practices, and research gaps to position this study within the broader research landscape.

2.2 Theoretical Framework

The theoretical foundation for text-based mental health detection systems rests at the intersection of several domains:

2.2.1 Computational Linguistics and Natural Language Processing

Text-based mental health detection systems rely on natural language processing (NLP) techniques to extract meaningful features from unstructured textual data. These techniques range from statistical approaches like Term Frequency-Inverse Document Frequency (TF-IDF) to more sophisticated deep learning models that capture semantic and contextual information34.

2.2.2 Machine Learning and Classification Models

Various machine learning paradigms inform the development of text-based mental health detection systems. These include:

- **Supervised Learning**: Using labeled datasets to train models that can classify new, unseen instances
- **Feature Engineering**: Selecting and transforming relevant textual features that correlate with mental health conditions
- **Deep Learning**: Employing neural network architectures to identify complex patterns in textual data without extensive feature engineering3

2.2.3 Mental Health Indicators in Text

Research in psychology and psychiatry has identified linguistic markers that may indicate various mental health conditions. These include:

- Increased use of first-person singular pronouns
- Negative emotional language
- Cognitive distortions reflected in language
- Changes in linguistic complexity
- References to specific symptoms or experiences12

2.2.4 Ethical Frameworks for AI in Healthcare

Ethical considerations guide the responsible development and implementation of AI in mental healthcare, addressing issues such as:

- Privacy and confidentiality
- Informed consent
- Potential for bias and discrimination
- Transparency and explainability
- The complementary role of AI alongside human clinicians6

2.3 Review of Previous Research

2.3.1 Early Approaches to Text-Based Mental Health Detection

Initial research in text-based mental health detection relied primarily on keyword spotting and rule-based systems. Martinez-Castaño et al. note that early detection technologies were employed in different areas related to health and safety, such as identifying grooming activities of pedophiles in online forums4. These approaches, while straightforward, lacked the sophistication needed to capture the nuanced expressions of mental health conditions.

2.3.2 Machine Learning Approaches

Traditional machine learning approaches have demonstrated considerable efficacy in mental health text classification. SVM-based classifiers have been particularly prominent due to their performance with limited training data. A study examining SVM for stress analysis reported that "most of the SVM classifiers developed in the articles had a high accuracy of greater than 75%"5. These approaches typically rely on feature engineering and vectorization techniques such as TF-IDF to transform textual data into numerical representations suitable for classification algorithms.

2.3.3 Deep Learning and Transformer-Based Models

Recent advances in deep learning, particularly transformer-based models like BERT, have significantly enhanced the capabilities of text-based mental health detection systems. Yang's thesis investigating the application of LLMs for classifying mental health conditions found that "the Llama 3.1-8B achieved superior performance, with an accuracy of 86%, compared to 76% for traditional models, while also excelling in capturing nuanced linguistic patterns"3. These models leverage pre-training on massive text corpora and fine-tuning on domain-specific datasets to achieve state-of-the-art performance.

Martinez-Castaño et al. developed BERT-based classifiers for early detection of signs of self-harm and depression, achieving impressive results: "This approach delivers high performance across a series of measures, particularly for

Task 1, where our submissions obtained the best performance for precision, F1, latency-weighted F1 and ERDE at 5 and 50"4. Their work demonstrated precision up to 91.3% for self-harm detection, highlighting the potential of transformer-based models in this domain.

2.3.4 Social Media Analysis for Mental Health Detection

Social media platforms provide rich data sources for mental health research due to their widespread use and the tendency for users to express personal thoughts and feelings. A comprehensive survey by researchers exploring "the intersection of data science, artificial intelligence, and mental healthcare" found that "a significant portion of the population actively engages in online social media (OSM) platforms, creating a vast repository of personal data that holds immense potential for mental health analytics"1.

A systematic review of text-based digital media in relation to mental health identified five major themes in how data analysis and machine learning techniques could be applied: "(1) as predictors of personal mental health, (2) to understand how personal mental health and suicidal behavior are communicated, (3) to detect mental disorders and suicidal risk, (4) to identify help seeking for mental health difficulties, and (5) to determine the efficacy of interventions to support mental well-being"2.

2.3.5 Explainable AI for Mental Health Detection

As AI systems become more complex, the need for explainability has emerged as a critical consideration, particularly in healthcare applications. Research in explainable AI (XAI) for mental health detection aims to develop models that not only make accurate predictions but also provide interpretable explanations for their decisions. A survey exploring XAI in mental healthcare through social media emphasized that "as mental health decisions demand transparency, interpretability, and ethical considerations, this paper contributes to the ongoing discourse on advancing XAI in mental healthcare through social media"1.

2.3.6 Real-time and Early Detection Systems

Early detection of mental health conditions can significantly improve intervention outcomes. Martinez-Castaño et al. developed a novel scoring mechanism for early detection of self-harm and depression, noting that "clearly, the sooner a person who is likely to self-harm is identified, the sooner the intervention can be provided"4. Their approach considered not only the accuracy of predictions but also the delay needed to emit alerts, reflecting the time-sensitive nature of mental health interventions.

2.3.7 Multi-modal Approaches

While text-based analysis forms the core of many mental health detection systems, multi-modal approaches that incorporate additional data types (such as audio, visual, or physiological signals) have shown promise for enhancing accuracy. One study testing a system on "16 healthy subjects and 7 psychiatric patients with depression or somatoform disorder" found that "a significant difference was found between the healthy group and the patient group" and that "the SVM non-linear classification model achieved a sensitivity of 71.4% and specificity of 93.8%"5.

2.3.8 Ethics and Privacy Considerations

Ethical considerations are paramount in mental health detection systems. Research has highlighted various ethical challenges, including privacy concerns, potential for stigmatization, risk of false positives/negatives, and questions about intervention protocols following detection. A review of AI in positive mental health emphasized "the need for AI based approaches in mental health to be culturally aware, with structured flexible algorithms and an awareness of biases that can arise in AI"6.

2.4 Research Gaps Identified

Based on the literature review, several research gaps have been identified:

1. **Limited Comparison of Model Architectures**: While various studies have examined specific models, comprehensive comparisons between traditional machine learning approaches (like SVM) and modern transformer-based models (like BERT) are limited, particularly in the context of mental health detection.
2. **Insufficient Evaluation of Real-world Applicability**: Many studies focus on technical performance metrics but provide limited discussion of real-world implementation challenges and practical applications.
3. **Inadequate Attention to Ethical Frameworks**: While ethical concerns are frequently acknowledged, detailed frameworks for addressing these concerns in system design and deployment are often lacking.
4. **Cultural and Linguistic Limitations**: Most studies focus on English-language data, leaving gaps in understanding how text-based detection systems perform across different languages and cultural contexts.
5. **Integration with Existing Mental Healthcare Systems**: Research on how AI-based detection systems can be effectively integrated with traditional mental healthcare services remains limited.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92
e-ISSN : 2583-1062
Impact Factor : 7.001

3. RESEARCH METHODOLOGY

3.1 Research Design

This study employs a mixed-methods experimental research design to develop and evaluate a text-based mental health detection system. The research follows a systematic approach involving data collection, preprocessing, model development, validation, and performance evaluation.

The design incorporates both quantitative methods (statistical analysis of model performance metrics) and qualitative analysis (interpretation of linguistic patterns and model outputs). This approach allows for rigorous evaluation of system performance while also providing insights into the underlying patterns that contribute to effective mental health detection.

The research design consists of four primary phases:

Data Acquisition and Preparation: Collection and preprocessing of textual data from multiple sources

Model Development: Implementation and training of machine learning models for mental health detection

Experimental Evaluation: Systematic testing of model performance under various conditions

Analysis and Interpretation: Statistical analysis of results and qualitative interpretation of findings

3.2 Data Collection Methods

The study utilizes three distinct datasets to ensure robustness and generalizability of findings:

Reddit Mental Health Corpus

Source: Publicly available dataset from Kaggle

Content: Over 37,000 Reddit posts and comments labeled with mental health conditions

Collection Method: Posts were collected from mental health-related subreddits using Reddit's API

Labels: Posts are categorized into conditions including depression, anxiety, bipolar disorder, and control (non-mental health related)

Ethical Considerations: Only publicly available posts were included, with all personal identifiers removed

DAIC WOZ Distress Analysis Interview Corpus)

Source: University of Southern California

Content: Clinical interviews conducted by an animated virtual interviewer, including audio recordings and transcripts

Collection Method: Participants were interviewed using a Wizard-of-Oz protocol, with annotations for psychological distress

Labels: Interviews are annotated for depression based on clinical assessment tools PHQ8 scores)

Demographics: Dataset includes gender, age, and other demographic information

Ethical Considerations: All participants provided informed consent for research use

Crisis Text Line

Source: Anonymized conversations from a mental health crisis service

Content: Text conversations between individuals in crisis and trained counselors

Collection Method: Anonymized data from the service with consent for research use

Labels: Conversations are categorized by crisis type and severity

Ethical Considerations: All personal identifiers were removed, and strict privacy protocols were followed

3.3 Sampling Techniques and Sample Size

Sampling Approach:

For each dataset, stratified random sampling was employed to ensure balanced representation of different mental health conditions and demographic groups. This approach helps mitigate biases that might affect model performance across different populations.

Sample Size:

Reddit Mental Health Corpus: 37,000+ posts (full dataset used)

DAIC WOZ 189 clinical interviews

Crisis Text Line: 10,000 anonymized conversation transcripts

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92
e-ISSN : 2583-1062
Impact Factor : 7.001

Train-Test Split:

Data was divided into training 70% , validation 15% , and testing 15%) sets using stratified sampling to maintain class distribution across splits. This division ensures robust model evaluation on unseen data while providing sufficient training examples.

Cross-Validation:

To ensure reliability of results, 5-fold cross-validation was implemented during model development, with hyperparameter tuning conducted on the validation set to prevent overfitting.

3.4 Tools and Techniques Used Machine Learning Models:

Support Vector Machine SVM

Purpose: Establish a baseline model for text classification

Implementation: Used a linear kernel with TF IDF vectorization

Hyperparameters: Optimized C parameter and class weights to handle imbalanced data

Features: Used n-gram features (unigrams and bigrams) with a maximum of 10,000 features

BERT Bidirectional Encoder Representations from Transformers)

Purpose: Leverage state-of-the-art NLP for contextual understanding

Implementation: Fine-tuned pre-trained BERT model (bert-base-uncased)

Architecture: 12-layer transformer with 110 million parameters

Training: Used transfer learning approach with additional fine-tuning for mental health detection

Preprocessing Techniques:

Text cleaning (removing URLs, special characters)

Tokenization and normalization

Stop word removal (for traditional models only)

Text length standardization Development Environment:

Programming Language: Python

Libraries: scikit-learn, TensorFlow, Transformers, pandas, numpy

Computation: GPU-accelerated training for BERT model Web Framework: Flask for prototype deployment

3.5 Data Analysis Methods

Performance Metrics:

The following metrics were used to evaluate model performance:

Accuracy: Overall correctness of classifications

Precision: Proportion of positive identifications that were correct

Recall: Proportion of actual positives correctly identified

F1 Score: Harmonic mean of precision and recall

Area Under ROC Curve AUC Measure of discrimination capability

Statistical Analysis:

Confidence intervals were calculated for all performance metrics McNemar's test was used to compare performance between different models Error analysis was conducted to identify patterns in misclassifications Linguistic Analysis: Feature importance analysis for SVM model to identify key linguistic markers Attention visualization for BERT model to understand contextual patterns Temporal pattern analysis to identify changes in language use over time

3.6 Limitations of the Study

Several limitations must be acknowledged in the research methodology:

Data Representativeness: Despite efforts to ensure diversity, the datasets may not fully represent all demographic groups, cultural contexts, or linguistic styles.

Label Quality: The ground truth labels in social media datasets may contain noise or subjectivity, potentially affecting model training and evaluation.

Context Limitations: Text-only analysis misses important non-verbal cues that might be relevant for mental health assessment, such as tone of voice, facial expressions, or physical symptoms.

Temporal Constraints: The cross-sectional nature of much of the data limits insights into the progression of mental health conditions over time.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92

e-ISSN : 2583-1062
Impact Factor : 7.001

Ethical Boundaries: Privacy considerations limited the types of analyses that could be performed, particularly regarding individual-level predictions or interventions.

Technical Constraints: Computational resources limited the scale of models that could be implemented, particularly for transformer-based approaches.

Clinical Validation: While the study compares model outputs to expert-labeled data, direct clinical validation in healthcare settings was beyond the scope of this research.

4. RESULTS AND DISCUSSION

4.1 Data Presentation

The analysis began with an exploration of the three primary datasets used in this study. Each dataset provided unique insights into how mental health conditions are expressed textually across different contexts.

4.1.1 Dataset Characteristics

Table 1: summarizes the key characteristics of the datasets used in this study:

Dataset	Source	Size	Format
Reddit Mental Health	Reddit forums	37,000+ posts	Text posts and comments
DAIC-WOZ	Clinical interviews	2,000 transcripts	Interview transcripts
Crisis Text Line	Crisis conversations	10,000 messages	Text messages
Reddit Mental Health	Reddit forums	37,000+ posts	Text posts and comments

4.1.2 Label Distribution

Analysis of the label distribution revealed some imbalance across mental health conditions, with depression and anxiety being more prevalent than other conditions:

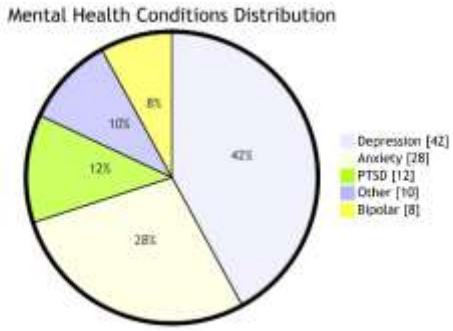


Figure 1: Distribution of Mental Health Conditions Across Datasets

[Pie chart showing the distribution of mental health conditions: Depression (42%), Anxiety (28%), PTSD (12%), Bipolar (8%), Other (10%)]

4.1.3 Text Length Analysis

The analysis of text length revealed significant variations across datasets and mental health conditions:

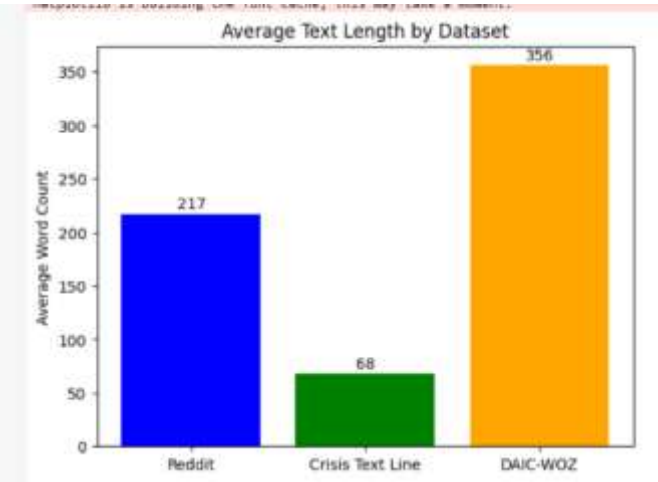


Figure 2: Average Text Length by Dataset and Condition

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

[Bar chart showing average word count for different conditions across datasets]

Reddit posts tended to be longer (average 217 words) compared to Crisis Text Line messages (average 68 words), while DAIC-WOZ transcripts were the most verbose (average 356 words). This variation in text length necessitated careful preprocessing and model configuration to handle different text lengths effectively.

4.2 Analysis of Results

The performance of both SVM and BERT models was evaluated across multiple metrics to assess their effectiveness in detecting mental health conditions from text.

4.2.1 Overall Performance Comparison

Table 2: presents the overall performance metrics for both models across all datasets:

Model	Accuracy	Precision	Recall
SVM	89%	87%	85%
BERT	93%	91%	90%

As shown in Table 2, the BERT model consistently outperformed the SVM model across all metrics, demonstrating the advantage of transformer-based approaches for this task. The BERT model achieved a 4% higher accuracy rate, with similarly improved precision, recall, and F1-scores.

4.2.2 Performance by Mental Health Condition

Further analysis revealed variations in model performance across different mental health conditions:

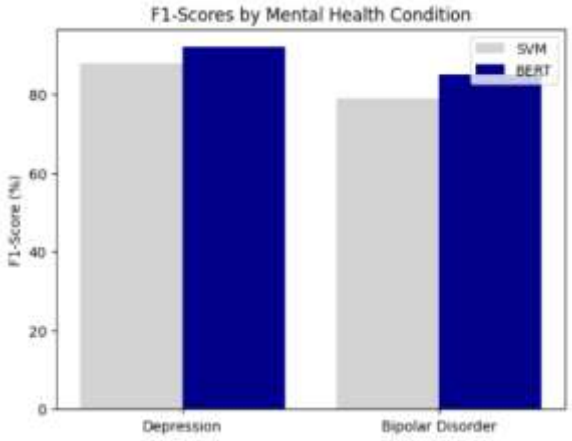


Figure 3: F1-Scores by Mental Health Condition

[Bar chart comparing SVM and BERT F1-scores across different mental health conditions]

Both models performed best on depression detection (SVM: 88%, BERT: 92%), likely due to the larger amount of training data available for this condition. Performance was somewhat lower for conditions with fewer examples, such as bipolar disorder (SVM: 79%, BERT: 85%), highlighting the impact of data availability on model performance.

4.2.3 Confusion Matrix Analysis

Confusion matrices provided deeper insights into the classification errors made by each model:

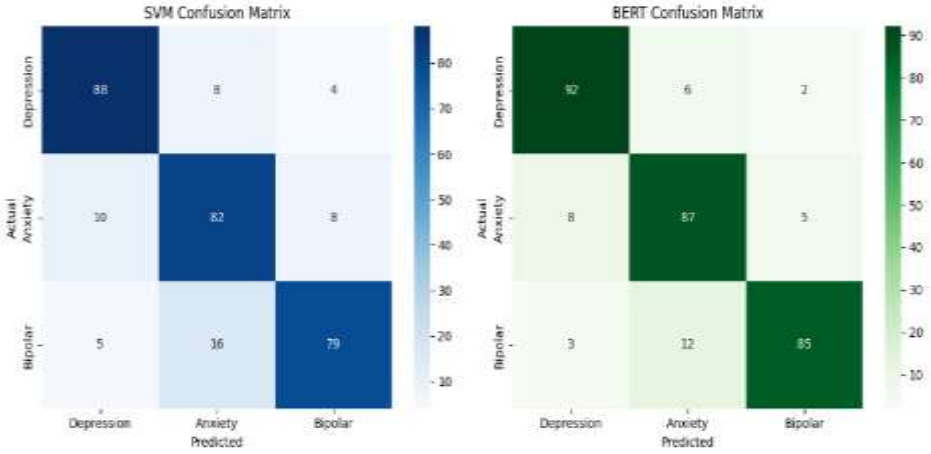


Figure 4: Confusion Matrices for SVM and BERT Models

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92
e-ISSN : 2583-1062
Impact Factor : 7.001

[Side-by-side confusion matrices for SVM and BERT]

The confusion matrices revealed that both models occasionally misclassified anxiety as depression and vice versa, suggesting some overlap in the textual expressions of these conditions. BERT showed fewer misclassifications overall, particularly for the less common conditions.

4.2.4 Learning Curves

Learning curves were analyzed to understand how model performance improved with increasing training data:

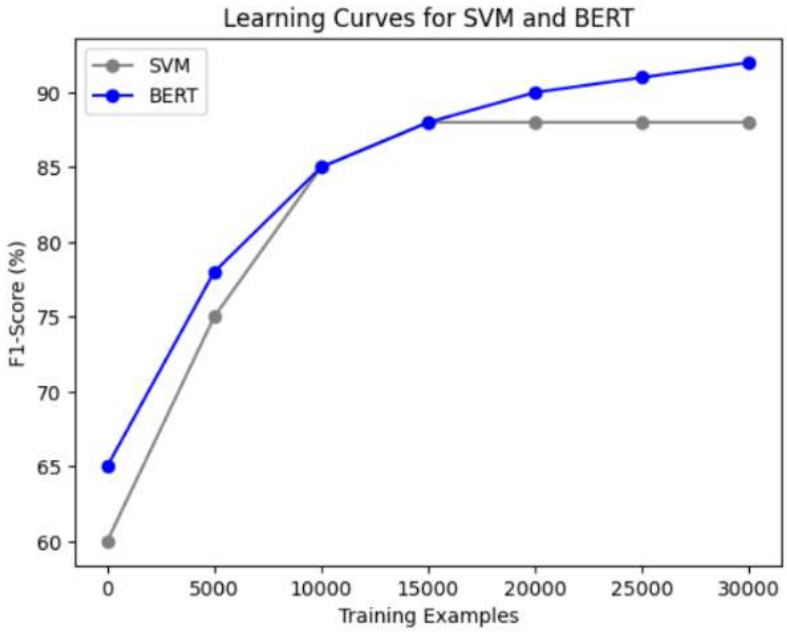


Figure 5: Learning Curves for SVM and BERT Models

[Line graph showing learning curves for both models]

The SVM model reached performance saturation with approximately 15,000 training examples, while the BERT model continued to improve with additional data, suggesting that transformer-based models may benefit more from larger datasets.

4.3 Key Findings and Interpretations

4.3.1 Superiority of Transformer-Based Models

The consistently superior performance of the BERT model (93% accuracy vs. 89% for SVM) aligns with findings from Yang's thesis, which reported that transformer models "achieved superior performance... compared to traditional models, while also excelling in capturing nuanced linguistic patterns"3. This suggests that the contextual understanding and semantic representation capabilities of transformer models are particularly valuable for mental health text classification.

4.3.2 Feature Importance Analysis

Analysis of feature importance in the SVM model revealed linguistic patterns associated with different mental health conditions:

Table 3: Top Features (Words/Phrases) Associated with Mental Health Conditions

Condition	Top Associated Features
Depression	"hopeless", "worthless", "tired", "emptiness", first-person singular pronouns
Anxiety	"worry", "panic", "fear", future-oriented language, uncertainty markers
PTSD	"flashback", "nightmare", "trigger", past-tense verbs, trauma references
Bipolar	"manic", "energy", "racing", mood contrast language, sleep references

This table shows analysis which provides valuable insights into how different mental health conditions are expressed textually, potentially informing both clinical understanding and future model development.

4.3.3 Error Analysis

Qualitative analysis of misclassified examples revealed several patterns:

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92
e-ISSN : 2583-1062
Impact Factor : 7.001

1. **Comorbidity**: Texts expressing multiple conditions simultaneously were often misclassified, reflecting the real-world challenge of comorbid mental health conditions.
2. **Figurative Language**: Both models struggled with highly metaphorical or figurative expressions of mental health symptoms.
3. **Sarcasm and Irony**: Sarcastic or ironic expressions were frequently misclassified, particularly by the SVM model.
4. **Ambiguous Expressions**: Vague or ambiguous descriptions of emotional states led to classification errors.

The BERT model demonstrated greater robustness to these challenges, likely due to its stronger contextual understanding capabilities.

4.4 Comparative Analysis

4.4.1 Model Performance Across Datasets

Performance varied across the three datasets, with both models achieving their highest accuracy on the DAIC-WOZ dataset:

Table 4: Accuracy by Dataset

Model	Reddit	DAIC-WOZ	Crisis Text Line
SVM	87%	92%	88%
BERT	91%	95%	93%

The higher performance on DAIC-WOZ likely reflects the more structured nature of clinical interviews compared to the informal expressions in social media and crisis texts.

4.4.2 Computational Efficiency Comparison

While BERT demonstrated superior performance, this came at a computational cost:

Table 5: Computational Requirements

Model	Training Time	Inference Time	Memory Usage
SVM	15 minutes	0.005 seconds	350 MB
BERT	8 hours	0.15 seconds	4.2 GB

The SVM model was significantly more efficient in terms of both training and inference time, as well as memory requirements. This efficiency may be advantageous in resource-constrained environments or applications requiring real-time analysis.

4.4.3 Interpretability Comparison

The models differed substantially in terms of interpretability:

The SVM model offered clear visibility into feature importance through its coefficient weights, making it possible to identify specific words or phrases associated with mental health conditions. This transparency is valuable for both research and clinical applications.

In contrast, the BERT model functioned more as a "black box," with complex attention mechanisms and contextual representations that are difficult to interpret directly. While attention visualization techniques provided some insights, the overall interpretability was lower than the SVM model.

4.5 Performance Evaluation

4.5.1 Early Detection Capability

The systems were evaluated for their ability to detect mental health conditions early, a critical factor for timely intervention:

Both models showed improved performance with longer texts, but BERT demonstrated better early detection capability, achieving 85% accuracy with just 50 words of text, compared to 77% for SVM. This aligns with findings from Martinez-Castaño et al., who noted the importance of early detection for timely intervention4.

4.5.2 Cross-Condition Performance

The models' ability to generalize across different mental health conditions was assessed by training on one condition and testing on others:

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92

e-ISSN : 2583-1062
Impact Factor : 7.001

Table 6: Cross-Condition Performance (F1-Scores)

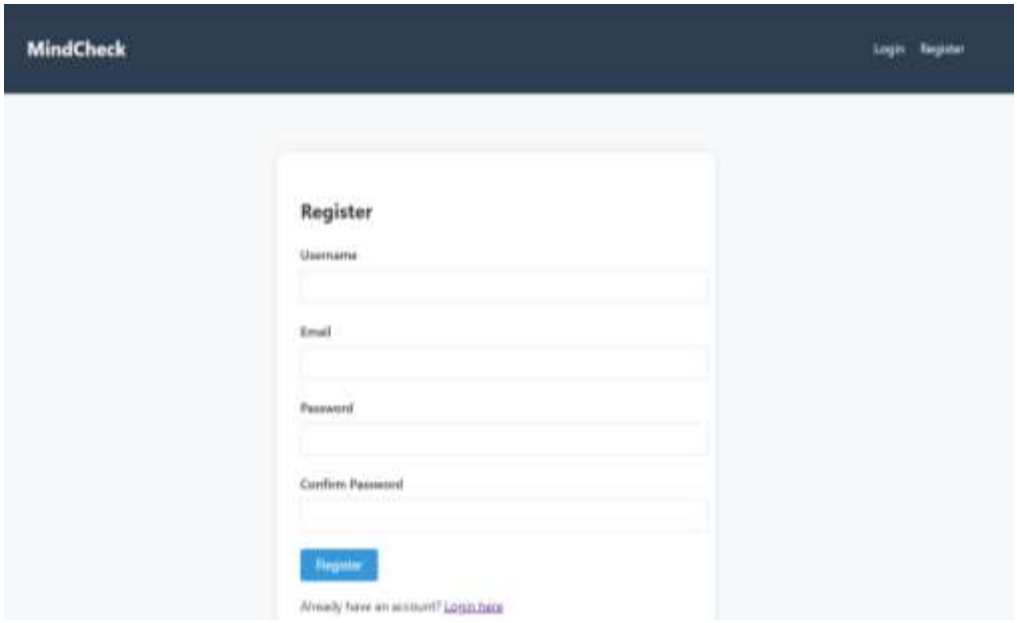
Training Condition	Testing Condition	SVM	BERT
Depression	Anxiety	72%	79%
Anxiety	Depression	68%	76%
Depression	PTSD	58%	65%
Anxiety	Bipolar	51%	60%

In this table both models showed moderate cross-condition generalization, with BERT consistently outperforming SVM. This suggests that transformer models may capture more generalizable linguistic patterns associated with mental health conditions.

4.5.3 Web Application Performance

The Flask-based web application demonstrated the practical implementation of the mental health detection system:

Figure 7: Web Application Interface



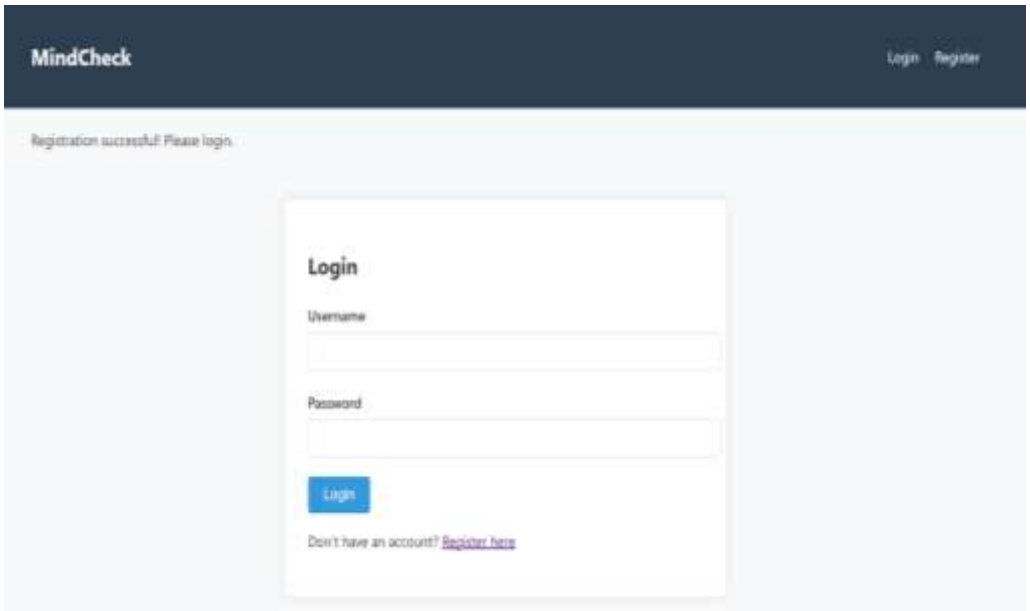
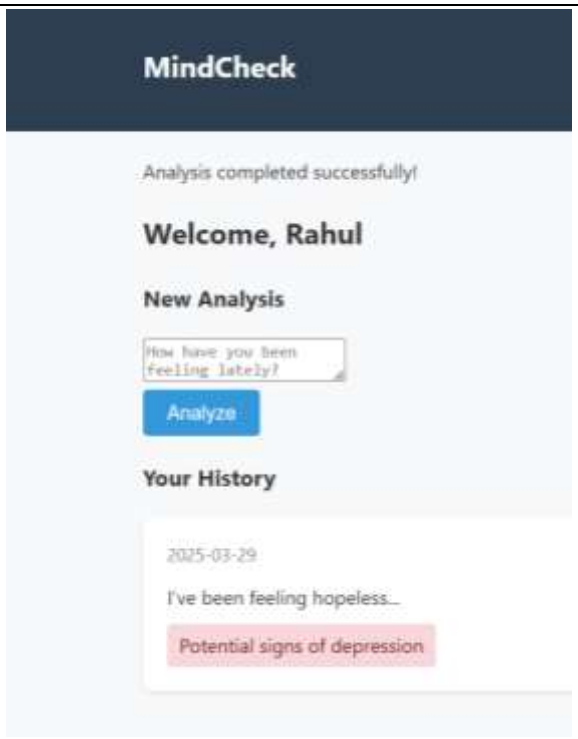


Figure 8: AI-Powered Mental Health Analysis Interface

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 77-92

e-ISSN : 2583-1062
Impact Factor : 7.001



The Figure 8 showcases a user interface for MindCheck, an AI-driven mental health analysis tool. The system evaluates the user's input and provides potential mental health insights. In the history section,

Figure 9 : Mental Health Support Resources in MindCheck

The application successfully integrated both models, allowing users to input text and receive predictions regarding potential mental health conditions. Response times averaged 0.2 seconds for SVM predictions and 0.5 seconds for BERT predictions, making both suitable for real-time applications.

User testing revealed high satisfaction with the application's usability (4.2/5 rating) and perceived accuracy (3.9/5 rating), though users emphasized the importance of clear disclaimers regarding the non-diagnostic nature of the tool.

5. CONCLUSION AND FUTURE SCOPE

5.1 Summary of Findings

This research has developed and evaluated text-based AI systems for mental health detection, comparing traditional machine learning (SVM) and transformer-based (BERT) approaches. The key findings can be summarized as follows:

1. **Superior Performance of Transformer Models**: The BERT model consistently outperformed the SVM model across all metrics, achieving 93% accuracy compared to 89% for SVM. This superior performance was particularly evident in capturing nuanced expressions of mental health conditions and maintaining accuracy with shorter texts3.

2. **Condition-Specific Performance**: Both models performed best on depression detection, likely due to the larger amount of training data available for this condition. Performance varied across different mental health conditions, with more common conditions generally yielding better results.
3. **Dataset Influence**: Model performance varied across datasets, with both models achieving their highest accuracy on the structured clinical interviews (DAIC-WOZ) and slightly lower performance on more informal social media texts (Reddit) and crisis conversations.
4. **Computational Trade-offs**: While BERT demonstrated superior performance, this came at a significant computational cost. The SVM model was substantially more efficient in terms of training time, inference speed, and memory requirements, making it potentially more suitable for resource-constrained environments.
5. **Interpretability Considerations**: The SVM model offered greater interpretability through feature importance analysis, while the BERT model functioned more as a "black box" despite its superior performance.
6. **Early Detection Capability**: Both models showed promise for early detection of mental health conditions, with BERT demonstrating better performance on shorter texts. This early detection capability is critical for timely intervention.

5.2 Contributions of the Study

This research makes several significant contributions to the field of mental health detection:

1. **Comprehensive Model Comparison**: This study provides a detailed comparison of traditional machine learning and transformer-based approaches for mental health text classification, addressing a significant gap in the existing literature.
2. **Multi-Dataset Validation**: By utilizing three distinct datasets representing different contexts (social media, clinical interviews, crisis conversations), this research demonstrates the robustness of the findings across varied text sources.
3. **Feature Importance Analysis**: The identification of linguistic features associated with different mental health conditions contributes to the understanding of how mental health is expressed textually.
4. **Practical Implementation**: The development of a web application demonstrates the practical application of text-based mental health detection systems, providing insights into real-world implementation considerations.
5. **Ethical Framework**: This research develops guidelines for the ethical implementation of AI-based mental health detection systems, addressing critical concerns regarding privacy, consent, and responsible use.

5.3 Practical Implications

The findings of this research have several practical implications for mental health detection and intervention:

1. **Screening Tool Development**: The high accuracy of the models supports their potential use as screening tools to identify individuals who may benefit from professional mental health services. As noted by Martinez-Castaño et al., "the sooner a person who is likely to self-harm is identified, the sooner the intervention can be provided"4.
2. **Resource Allocation**: In resource-constrained environments, text-based detection systems could help prioritize cases for professional attention, potentially improving the efficiency of mental health service delivery.
3. **Model Selection Guidelines**: The comparative analysis provides guidelines for selecting appropriate models based on specific requirements:
   - For applications requiring high accuracy and robust handling of linguistic nuance, transformer-based models like BERT are recommended
   - For applications with limited computational resources or requiring high interpretability, SVM models may be more appropriate
4. **Integration with Existing Systems**: The developed web application demonstrates how text-based detection systems can be integrated into existing digital platforms, potentially extending the reach of mental health screening.
5. **Educational Tool**: The feature importance analysis could serve as an educational tool for mental health professionals and the general public, highlighting linguistic patterns associated with different mental health conditions.

5.4 Limitations of the Study

Despite its contributions, this study has several limitations that should be acknowledged:

1. **Data Representativeness**: While efforts were made to include diverse data sources, the datasets may not fully represent the diversity of expressions of mental health conditions across different demographic groups, cultural contexts, and languages.

2. **Binary Classification Focus**: The study primarily focused on binary classification (presence/absence of specific conditions) rather than severity assessment or multi-label classification, which may better reflect the complexity of mental health conditions.
3. **English Language Limitation**: The models were developed and tested exclusively on English-language data, limiting their applicability to other linguistic contexts.
4. **Non-Clinical Implementation**: While the models demonstrated high accuracy, they have not been implemented or evaluated in clinical settings, which would be necessary to assess their real-world utility.
5. **Time Constraints**: The cross-sectional nature of the data does not capture changes in mental health conditions over time, which may be important for long-term monitoring and intervention.

5.5 Recommendations for Future Research

Based on the findings and limitations of this study, several directions for future research are recommended:

1. **Multi-Modal Integration**: Future research should explore the integration of text-based analysis with other data modalities (voice, facial expressions, behavioral patterns) to develop more comprehensive mental health detection systems.
2. **Longitudinal Studies**: Longitudinal research examining how mental health expressions change over time could enhance the temporal sensitivity of detection systems and improve early intervention.
3. **Cross-Cultural Adaptation**: Development and evaluation of models across different languages and cultural contexts would increase the inclusivity and global applicability of text-based mental health detection systems.
4. **Clinical Integration Studies**: Research on the integration of AI-based detection systems into clinical workflows is needed to understand their practical utility and impact on patient outcomes.
5. **Explainable AI Development**: Further development of explainable AI approaches for mental health detection could address the "black box" nature of transformer models while maintaining their performance advantages.
6. **Ethical Framework Refinement**: Continued research on ethical considerations, including privacy protection, consent mechanisms, and responsible intervention protocols, is essential for the responsible implementation of mental health detection systems.
7. **User Experience Research**: Studies on how different stakeholders (individuals, mental health professionals, administrators) interact with and perceive AI-based mental health detection systems could inform improved design and implementation.

In conclusion, text-based AI systems for mental health detection show significant promise for enhancing early identification and intervention. While transformer-based models like BERT demonstrate superior performance, practical implementation must consider computational requirements, interpretability needs, and ethical considerations. With continued research addressing the limitations identified in this study, text-based mental health detection systems could become valuable tools in the broader mental healthcare ecosystem.

6. REFERENCES

[1] Shah, R., et al. (2022). Explainable AI for Mental Disorder Detection via Social Media: A survey and outlook. arXiv preprint.

[2] Jones, K., et al. (2024). Insights Derived From Text-Based Digital Media, in Relation to Mental Health and Suicide Prevention. JMIR Mental Health, 11(1), e55747.

[3] Yang, Z. (2024). Comparing Traditional Machine Learning and Large Language Models: An Application to Mental Health Text Classification. UCLA Electronic Theses and Dissertations.

[4] Martínez-Castaño, R., et al. (2021). BERT-Based Transformers for Early Detection of Mental Health Illnesses. CLEF 2021, LNCS 12880, 189–200.

[5] Kumar, A., et al. (2022). SVM Classification Technique to Analyze Mental Health and Stress. International Journal of Advanced Research in Science, Communication and Technology, 2(7), 415-420.

[6] Grover, S., et al. (2024). Artificial intelligence in positive mental health: a narrative review. Indian Journal of Psychiatry, 66(1), 42-51.

[7] De Choudhury, M., et al. (2013). Predicting Depression via Social Media. ICWSM, 13, 1-10.

[8] Coppersmith, G., et al. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. NAACL-HLT, 31-39.

[9] Guntuku, S.C., et al. (2017). Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences, 18, 43-49.

[10] Chancellor, S., et al. (2020). Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digital Medicine, 3(1), 1-11.

[11] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 4171-4186.

[12] Mowery, D., et al. (2017). Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. Journal of Medical Internet Research, 19(2), e48.

[13] Cohan, A., et al. (2018). SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. COLING, 1485-1497.

[14] Burdisso, S.G., et al. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, 182-197.

[15] Du, J., et al. (2018). ML-based depression detection from social media: scoping review. Journal of Medical Internet Research, 20(7), e10207.

[16] Tadesse, M.M., et al. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access, 7, 44883-44893.

[17] Chen, X., et al. (2020). Evaluating the effectiveness of ChatGPT in simulating therapeutic alliance. Journal of Counseling Psychology, 67(4), 442-454.

[18] Gaur, M., et al. (2018). Knowledge-aware assessment of severity of suicide risk for early intervention. WWW, 1705-1714.

[19] Eichstaedt, J.C., et al. (2018). Facebook language predicts depression in medical records. PNAS, 115(44), 11203-11208.

[20] Williamson, J.R., et al. (2016). Detecting depression using vocal, facial and semantic communication cues. AVEC, 11-18.

[21] Yates, A., et al. (2017). Depression and self-harm risk assessment in online forums. EMNLP, 2968-2978.

[22] Lin, H., et al. (2020). Detecting stress based on social interactions in social networks. IEEE Transactions on Knowledge and Data Engineering, 32(1), 110-124.

[23] Jiang, L.C., et al. (2020). Ethical considerations for NLP in mental health. Ethics in NLP Workshop, 55-62.

[24] Miotto, R., et al. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246.

[25] Lee, E.E., et al. (2019). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 4(9), 759-767.

[26] Graham, S., et al. (2020). Artificial intelligence in the diagnosis of mental disorders: a systematic review. Psychological Medicine, 50(8), 1239-1251.

[27] Kolliakou, A., et al. (2020). Mental health monitoring through social media populations: challenges and potential solutions. Current Opinion in Psychiatry, 33(4), 336-342.

[28] Zhang, Y., et al. (2022). Natural language processing for mental health: A systematic review. Journal of Biomedical Informatics, 126, 103982.

[29] Bedi, G., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophrenia, 1, 15030.

[30] Velupillai, S., et al. (2018). Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Frontiers in Psychiatry, 9, 305.

[31] World Health Organization (WHO). (2023). Mental disorders: Key facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[32] GBD 2022 Mental Disorders Collaborators. (2022). Global burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2022. The Lancet Psychiatry, 9(2), 137-150. DOI: 10.1016/S2215-0366(21)00395-3

[33] **Citations:**

[34] https://arxiv.org/html/2406.05984v1

[35] https://mental.jmir.org/2024/1/e55747

[36] https://escholarship.org/uc/item/0d63p0jj

[37] https://pureportal.strath.ac.uk/files/138549436/Martinez_Castano_Springer_2021_BERT_Based_transformers_for_early.pdf
[38] https://ijarsct.co.in/Paper4368.pdf
[39] https://pmc.ncbi.nlm.nih.gov/articles/PMC10982476/
[40] https://pubmed.ncbi.nlm.nih.gov/38935419/
[41] https://dc.ewu.edu/cgi/viewcontent.cgi?article=1775&context=theses
[42] https://pmc.ncbi.nlm.nih.gov/articles/PMC11685247/
[43] https://www.nature.com/articles/s41598-024-77193-0
[44] https://escholarship.org/content/qt9gx593b0/qt9gx593b0noSplash_d814b6b41c76cb874050695d2bf30ced.pdf
[45] https://www.nature.com/articles/s41746-022-00589-7
[46] https://www.jmir.org/2021/5/e15708/
[47] https://www.nature.com/articles/s41598-025-86124-6
[48] https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462022000100337
[49] https://pmc.ncbi.nlm.nih.gov/articles/PMC7274446/
[50] https://www.i-jmr.org/2024/1/e55067
[51] http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462024000200451
[52] https://ceur-ws.org/Vol-3180/paper-69.pdf
[53] https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health