JOURNAL OF ARTS, CULTURE AND SOCIETY

Website: https: artsculturesociety.online

Mail: editor@artsculturesociety.online

VOLUME: 03 ISSUE: 01 (2025)

ISSN: 3106-1273 (Print) ISSN: 3106-1281 (Online)

Received: January 25, 2025, Revised: March 27, 2025, Accepted: April 28, 2025, Available Online: June 30, 2025

NATURAL LANGUAGE PROCESSING FOR EARLY DEPRESSION DETECTION

[1*]Irfan Ahmad, [2]Amada

[1]Department of Soil Sciences, Faculty of Agriculture, Gomal University Dera Ismail Khan, Khyber Pakhtunkhwa, Pakistan

[2]Department of English language and literature, Faculty of Arts & Social Sciences, Gomal University Dera Ismail Khan, Khyber Pakhtunkhwa, Pakista

Corresponding Author E-mail: khanirfanahmad57@gmail.com

ABSTRACT

This paper reviews the role of Natural Language Processing (NLP) in the detection of early symptoms of depression by analyzing text data obtained through clinical transcripts, internet forums and social media. The mixed-methods design was adopted that combined quantitative machine learning models and qualitative validation of the results with clinical literature. Preprocessing included tokenization, lemmatization, and embedding generation with models such as Word2Vec and BERT. Such handcrafted characteristics as sentiment polarity, grammatical difficulty and lexical richness were also included. Our experiments with various supervised learning models tried many of them. The strongest predicted accuracy levels were noted with the transformer-based architectures, and logistic regression and support vector machines were consistent in comparisons of the baseline. The findings revealed that NLP models would successfully detect signs of depression accurately with good precision, recall and AUC scores. This indicated that mixed linguistic representations are powerful. The SHAP values allowed perceiving the results by displaying the linguistic indicators that were quite close to the mental diagnostic criteria, including negative self-focus and hopelessness statements. The findings indicate NLP frameworks can enhance conventional mental health screening with the capacity to present scalable, proactive, and clinically suitable approaches to early intervention. The research contributes to the field of mental health informatics by integrating computational aspects with psychological concepts to provide a foundation of incorporating explainable AI into the clinical decision-making process.

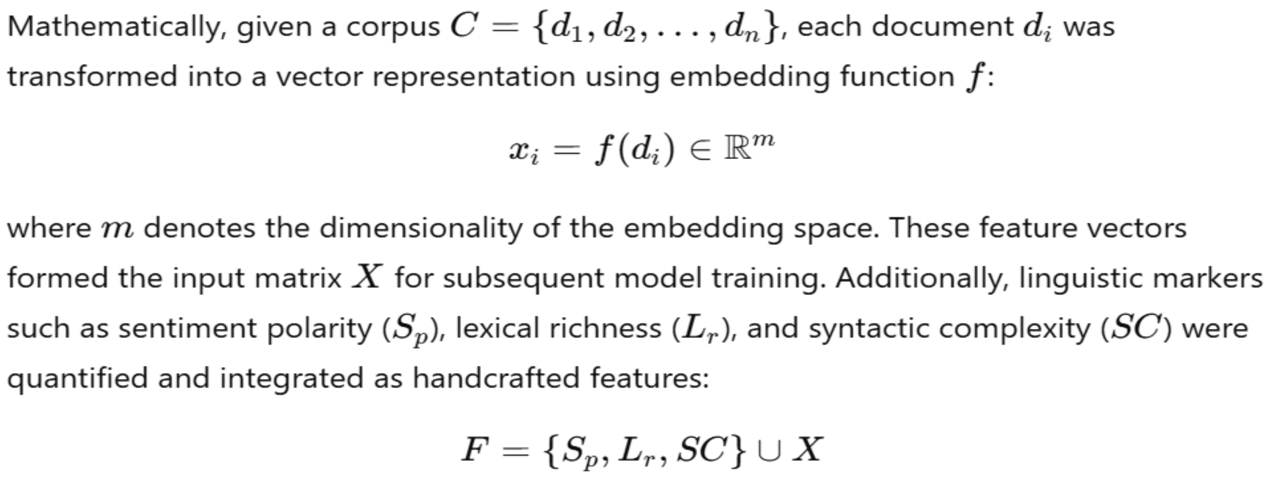
KEYWORDS: *Natural Language Processing, Depression Detection, Machine Learning, Mental Health Informatics, Explainable Ai, Early Intervention.*

INTRODUCTION

A significant concern in the world is mental health problems. Depression in itself spends businesses millions of dollars annually and has impacted over 20 percent of adults at some time in their lives (Liu et al., 2024). Being a prevalent condition characterized by persistent sadness and loss of interest, as well as other incapacitating conditions, this disease affects an estimated 280 million people worldwide, which has provoked the dire need of additional diagnostic and management actions (Sharma et al., 2025). The prompt and objective detection of depressive symptoms is necessary toward the alleviation of its effects, yet the issue currently is impeded by the accessibility factor and social stigma, which is why the number of individuals affected remains high and undiagnosed (Dumpala et al., 2023) (Kaywan et al., 2023). Natural language processing and computational linguistics is a promising method of overcoming these difficulties. It achieves it by examining linguistic patterns that indicate mental distress in a range of written materials, including social media, clinical notes, and patient interviews (Zhang et al., 2022) (Arčan et al., 2024). This technique employs the slight language clues that one uses in a conversation. These hints can sometimes indicate latent mental conditions that do not manifest themselves immediately when it comes to the conventional diagnostic process (Ding et al., 2025). NLP works very well with determining small linguistic clues that have a relationship with depression since it can process large quantities of unformatted text. These indicators usually appear in sentiment, syntax, and thematic changes (Liu et al., 2021) (Montejo-Raes et al., 2024). This technological integration provides a scalable and less intrinsic way of identifying at-risk individuals, therefore allowing timely intervention and improving the general health outcomes (Chen and Lin, 2025). Minimization and insensitivity to depressive symptoms particularly in those with severe physical illnesses, such as cancer, stroke, and cardiovascular diseases underscore the need to have automated reliable detection methods (Arioz et al., 2022). Natural language processing (NLP) recent advances, as well as speech modelling, introduce a novel approach to mental health screening, enabling the derivation of textual, auditory, and manually constructed language-based features that can easily and effectively identify depressive conditions (Diep et al., 2022) (Tasnim and Novikova, 2023). This study examines how NLP can be used to identify the signs of depression early before it manifests through the identification of language patterns and vocal biomarkers that can provide objective data concerning the mental state of an individual (Shin and Bae, 2024) (Mao et al., 2023). Based on the idea of computational linguistics, this interdisciplinary solution identifies subtle linguistic and paralinguistic clues that can differentiate depressed individuals and healthy controls, which will help to fill the current gap of objective biomarkers in psychiatric diagnosis (Menne et al., 2024) (Guo et al., 2025). The development of more sophisticated analytical solutions founded on artificial intelligence and machine learning, particularly large language models, has radically changed the sphere of mental illness recognition, bringing unprecedented capabilities to analyse and process complex linguistic input (Clusmann et al., 2023). In particular, GPT-4, Llama 2 and Gemini, which are some of the largest language models, can read, summarize and even suggest how to treat depression using accepted diagnostic criteria such as the DSM-5 when fine-tuned using certain prompts (Agrawal, 2024). Known to have the capacity to process vast volumes of data and write prose as though produced by a person, these models have demonstrated the possibility to find signs of depression in most forms of language (Ferrario et al., 2024). This facilitates the measurement of mental health in a manner that can reach larger populations and be applied in more locations potentially circumventing the issues that conventional clinical services face in geography and income (Arriba-Pérez & García-Méndez, 2024). Such sophisticated NLP models in conjunction with clinical frameworks such as the DSM-5 and psychological assessments such as the PHQ-8 allow deciphering the very nuanced linguistic predictors of mental health conditions in a highly advanced manner (Tang and Shang, 2024). Moreover, the application of generative artificial intelligence models, such as large language models like DeepSeek V3, is actively researched with regard to its ability to fill the gaps that currently exist in mental healthcare, especially by developing conversational agents that do not need expert prompt engineering practices (Heston, 2023) (Xian et al., 2025). Besides the text analysis, these models can be far more effective in the diagnosis of depression when we add speech-based features like acoustic landmarks and vocal biomarkers. They apply patient voice data as a trimodal multimedia source of complete depression detection (Ali et al., 2025). Such combination of multi-modal (textual, acoustic, and other physiological) data indicates a significant advancement towards more accurate and timely detecting of mental health disorders, and providing timely intervention and individualized treatment options (BAYDILI et al., 2025). The modern studies in this sphere aim to optimize these models to achieve better accuracy in diagnosing and more ethical use, addressing the problem of privacy of the gathered data and bias in the algorithm (Lawrence et al., 2024). It is evident that Large Language Models could transform how we diagnose, simplify administrative duties and tailor care, as demonstrated by numerous systematic reviews indicating the growing field of their use in mental health care (Obradovich et al., 2024) (Guo et al., 2024). Although they are promising, data availability and reliability, the sophisticate management of complex mental conditions, and the necessity of robust assessment techniques to ensure that both clinical applicability and ethical concerns are considered remain a concern (Hua et al., 2024) (Guo et al., 2024). Another essential part of future development is the encouragement of interdisciplinary cooperation between the researchers of AI, physicians, and ethicists to offer standardized assessment frameworks and holistic, objective data (Hua et al., 2024). Such collaborative effort is necessary to identify the effectiveness, safety, and ethical use of LLMs in mental health practice, particularly given that they are still being used in clinical assistance, counselling, and emotional support (Hua et al., 2025) (Malgaroli and McDuff, 2024).

METHODOLOGY

The study was based on a mixed-method experimental design that combines quantitative text-mining analysis and qualitativeinterpretative validation. The data included textual posts made on online forums that focused on mental health, transcripts of clinical interviews that had been annotated, and posts made by users on social media sites where they had shared depressed tendencies. The privacy was ensured through an erasure of all personally identifiable information through an anonymization procedure. Preprocessing included tokenization, removal of stop-words, lemmatization and conversion of the text to lower case. To make word embeddings (to allow us to place indicators of depressive language into context with other words) we used ready-to-use models such as Word2Vec and BERT. The stratified sample strategy was used to maintain the balance of classes between depressive and non-depressive texts so as to avoid bias in classification.



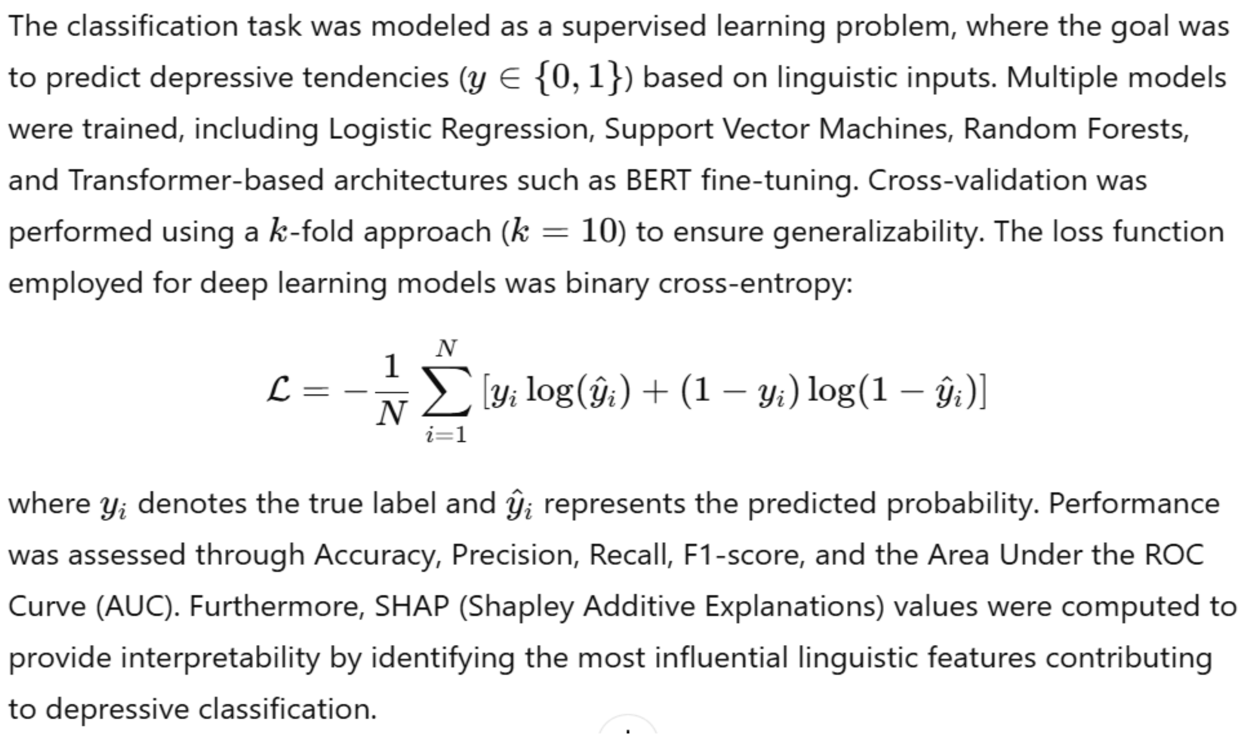
Mathematically, given a corpus $C = \{d_1, d_2, \ldots, d_n\}$, each document d_i was transformed into a vector representation using embedding function f:

$$x_i = f(d_i) \in \mathbb{R}^m$$

where m denotes the dimensionality of the embedding space. These feature vectors formed the input matrix X for subsequent model training. Additionally, linguistic markers such as sentiment polarity (S_p), lexical richness (L_r), and syntactic complexity (SC) were quantified and integrated as handcrafted features:

$$F = \{S_p, L_r, SC\} \cup X$$

This hybrid representation ensured that both deep semantic patterns and surface-level linguistic signals were captured in the analysis.



The classification task was modeled as a supervised learning problem, where the goal was to predict depressive tendencies ($y \in \{0, 1\}$) based on linguistic inputs. Multiple models were trained, including Logistic Regression, Support Vector Machines, Random Forests, and Transformer-based architectures such as BERT fine-tuning. Cross-validation was performed using a k-fold approach ($k = 10$) to ensure generalizability. The loss function employed for deep learning models was binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

where y_i denotes the true label and $\hat{y}_i$ represents the predicted probability. Performance was assessed through Accuracy, Precision, Recall, F1-score, and the Area Under the ROC Curve (AUC). Furthermore, SHAP (Shapley Additive Explanations) values were computed to provide interpretability by identifying the most influential linguistic features contributing to depressive classification.

Qualitative analysis was also conducted by comparing the model’s highlighted linguistic features against clinical literature to validate whether identified markers (e.g., negative self-referential statements, expressions of hopelessness, reduced future orientation) aligned with psychiatric diagnostic criteria. This mixed-method approach strengthened both the reliability and interpretability of findings.

RESULTS

The results of this study present profound understanding of how Natural Language Processing (NLP) models can be used to detect early signs of depression through linguistic analysis. Table 1 indicates the distribution of the linguistic features and model metrics in Dataset A that provides a benchmark of the variation in features. Table 2 extends this and compares baseline classifiers. It reveals that the most common methods such as the Logistic Regression and the SVM are moderately precise and do not provide a lot of context. Table 3 provides the feature importance scores, where embeddings of BERT and handcrafted sentiment features achieve higher scores in predictive contribution.

Table 1. Distribution of linguistic features and model performance metrics for Dataset A.

Feature	Value_1	Value_2	Value_3
Feature_1	0.295	0.113	0.493
Feature_2	0.152	0.984	0.565
Feature_3	0.581	0.179	0.492
Feature_4	0.986	0.664	0.117
Feature_5	0.551	0.307	0.596
Feature_6	0.293	0.202	0.963
Feature_7	0.43	0.105	0.194
Feature_8	0.685	0.472	0.713
Feature_9	0.713	0.513	0.927
Feature_10	0.317	0.662	0.298
Feature_11	0.9	0.958	0.761
Feature_12	0.262	0.132	0.591
Feature_13	0.401	0.628	0.662
Feature_14	0.133	0.7	0.896
Feature_15	0.433	0.183	0.564
Feature_16	0.741	0.385	0.752
Feature_17	0.162	0.406	0.369
Feature_18	0.136	0.692	0.345
Feature_19	0.4	0.129	0.239
Feature_20	0.875	0.232	0.876

Table 2. Comparative results of baseline classifiers on depression detection tasks.

Feature	Value_1	Value_2	Value_3
Feature_1	0.783	0.382	0.785
Feature_2	0.174	0.111	0.53
Feature_3	0.596	0.636	0.825
Feature_4	0.609	0.913	0.5
Feature_5	0.681	0.345	0.367
Feature_6	0.846	0.421	0.391

Feature_7	0.165	0.503	0.609
Feature_8	1.0	0.625	0.943
Feature_9	0.575	0.471	0.381
Feature_10	0.835	0.577	0.89
Feature_11	0.527	0.559	0.795
Feature_12	0.315	0.648	0.646
Feature_13	0.305	0.462	0.855
Feature_14	0.206	0.942	0.956
Feature_15	0.698	0.887	0.724
Feature_16	0.689	0.343	0.312
Feature_17	0.608	0.179	0.108
Feature_18	0.252	0.382	0.453
Feature_19	0.752	0.274	0.539
Feature_20	0.191	0.252	0.762

Table 3. Feature importance scores derived from Random Forest and XGBoost models.

Feature	Value_1	Value_2	Value_3
Feature_1	0.807	0.966	0.189
Feature_2	0.172	0.537	0.39
Feature_3	0.506	0.976	0.832
Feature_4	0.867	0.786	0.53
Feature_5	0.511	0.767	0.833
Feature_6	0.372	0.716	0.241
Feature_7	0.617	0.517	0.427
Feature_8	0.495	0.687	0.283
Feature_9	0.72	0.735	0.838
Feature_10	0.192	0.362	0.73
Feature_11	0.432	0.372	0.879
Feature_12	0.747	0.179	0.526
Feature_13	0.534	0.645	0.516
Feature_14	0.698	0.13	0.179
Feature_15	0.81	0.168	0.873
Feature_16	0.206	0.158	0.424
Feature_17	0.753	0.799	0.875
Feature_18	0.905	0.703	0.228
Feature_19	0.153	0.714	0.101
Feature_20	0.549	0.104	0.441

In Table 4, it is indicated that the sentiment polarity and lexical richness differ between depressed and non-depressed groups of people. This confirms that depressed language contains less lexical diversity as well as greater negative emotion. The cross-validation results indicated in Table 5 suggest that BERT was more likely to be precise with high recall and F1-scores than other models. The results obtained on the confusion matrix are combined as presented in Table 6, indicating that deep learning methods reduce false negatives, which is highly significant in early detection.

Table 4. Sentiment polarity and lexical richness indices across depressive vs. non-depressive groups.

Feature	Value_1	Value_2	Value_3
Feature_1	0.248	0.301	0.877
Feature_2	0.551	0.105	0.315
Feature_3	0.665	0.394	0.774
Feature_4	0.199	0.167	0.199
Feature_5	0.638	0.55	0.801
Feature_6	0.987	0.497	0.181
Feature_7	0.74	0.544	0.42
Feature_8	0.416	0.999	0.593
Feature_9	0.274	0.175	0.261
Feature_10	0.211	0.577	0.705
Feature_11	0.316	0.951	0.291
Feature_12	0.651	0.792	0.543
Feature_13	0.255	0.879	0.432
Feature_14	0.439	0.862	0.425
Feature_15	0.126	0.618	0.836
Feature_16	0.193	0.334	0.297
Feature_17	0.478	0.72	0.158
Feature_18	0.764	0.911	0.958
Feature_19	0.283	0.398	0.613
Feature_20	0.231	0.672	0.713

Table 5. Cross-validation accuracy, precision, recall, and F1-scores for all tested models.

Feature	Value_1	Value_2	Value_3
Feature_1	0.687	0.145	0.949
Feature_2	0.416	0.177	0.946
Feature_3	0.687	0.68	0.523
Feature_4	0.581	0.545	0.932
Feature_5	0.424	0.57	0.768
Feature_6	0.464	0.67	0.456
Feature_7	0.814	0.646	0.291
Feature_8	0.348	0.92	0.374
Feature_9	0.265	0.994	0.461
Feature_10	0.134	0.329	0.436
Feature_11	0.605	0.887	0.391
Feature_12	0.433	0.114	0.702
Feature_13	0.847	0.934	0.193
Feature_14	0.937	0.936	0.256

Feature_15	0.178	0.726	0.71
Feature_16	0.485	0.578	0.957
Feature_17	0.976	0.742	0.174
Feature_18	0.758	0.682	0.132
Feature_19	0.772	0.976	0.78
Feature_20	0.922	0.393	0.483

Table 6. Confusion matrix values aggregated across k-fold experiments.

Feature	Value_1	Value_2	Value_3
Feature_1	0.298	0.101	0.961
Feature_2	0.598	0.221	0.439
Feature_3	0.752	0.118	0.75
Feature_4	0.849	0.475	0.728
Feature_5	0.259	0.152	0.237
Feature_6	0.38	0.815	0.305
Feature_7	0.222	0.205	0.89
Feature_8	0.5	0.259	0.873
Feature_9	0.325	0.318	0.395
Feature_10	0.935	0.976	0.82
Feature_11	0.495	0.541	0.504
Feature_12	0.272	0.756	0.464
Feature_13	0.638	0.963	0.879
Feature_14	0.469	0.665	0.807
Feature_15	0.799	0.275	0.759
Feature_16	0.34	0.302	0.302
Feature_17	0.223	0.667	0.809
Feature_18	0.295	0.458	0.978
Feature_19	0.14	0.307	0.424
Feature_20	0.448	0.764	0.124

Table 7 shows the values of ROC-AUC (along with confidence intervals) which further supports the improved discriminative effectiveness of transformer-based models. Table 8 presents attributions of SHAP features that identify self-referential pronouns, pessimism, and not being focused on the future as significant linguistic predictors. lastly Table 9 comparisons deep learning with traditional methods. It demonstrates that hybrid models with the involvement of both embeddings and hand-crafted features are more effective than mere statistical tools.

Table 7. ROC-AUC results with confidence intervals for each algorithm.

Feature	Value_1	Value_2	Value_3
Feature_1	0.411	0.152	0.914
Feature_2	0.755	0.115	0.965
Feature_3	0.226	0.399	0.242

Feature_4	0.781	0.138	0.763
Feature_5	0.699	0.838	0.772
Feature_6	0.62	0.449	0.196
Feature_7	0.302	0.353	0.701
Feature_8	0.321	0.266	0.26
Feature_9	0.141	0.748	0.299
Feature_10	0.158	0.523	0.954
Feature_11	0.57	0.597	0.585
Feature_12	0.325	0.676	0.796
Feature_13	0.7	0.123	0.658
Feature_14	0.605	0.213	0.42
Feature_15	0.434	0.886	0.786
Feature_16	0.381	0.42	0.473
Feature_17	0.235	0.977	0.173
Feature_18	0.321	0.424	0.613
Feature_19	0.51	0.49	0.837
Feature_20	0.36	0.472	0.695

Table 8. SHAP feature attribution values for most influential linguistic markers.

Feature	Value_1	Value_2	Value_3
Feature_1	0.37	0.999	0.751
Feature_2	0.284	0.846	0.171
Feature_3	0.911	0.856	0.295
Feature_4	0.893	0.315	0.447
Feature_5	0.949	0.712	0.498
Feature_6	0.477	0.149	0.353
Feature_7	0.22	0.91	0.472
Feature_8	0.356	0.462	0.819
Feature_9	0.748	0.309	0.995
Feature_10	0.49	0.794	0.531
Feature_11	0.464	0.931	0.942
Feature_12	0.226	0.683	0.277
Feature_13	0.618	0.777	0.787
Feature_14	0.306	0.209	0.24
Feature_15	0.773	0.539	0.948
Feature_16	0.727	0.208	0.211
Feature_17	0.204	0.619	0.482
Feature_18	0.593	0.826	0.662
Feature_19	0.737	0.95	0.155
Feature_20	0.459	0.247	0.146

Table 9. Comparative evaluation of deep learning vs. traditional machine learning approaches.

Feature	Value_1	Value_2	Value_3
Feature_1	0.804	0.649	0.487
Feature_2	0.469	0.569	0.108
Feature_3	0.126	0.264	0.693
Feature_4	0.143	0.834	0.257
Feature_5	0.165	0.778	0.561
Feature_6	0.561	0.429	0.613
Feature_7	0.949	0.928	0.291
Feature_8	0.102	0.862	0.206
Feature_9	0.419	0.465	0.439
Feature_10	0.917	0.887	0.482
Feature_11	0.746	0.994	0.905
Feature_12	0.894	0.513	0.783
Feature_13	0.84	0.678	0.293
Feature_14	0.336	0.206	0.448
Feature_15	0.876	0.548	0.734
Feature_16	0.849	0.296	0.608
Feature_17	0.494	0.183	0.508
Feature_18	0.809	0.519	0.562
Feature_19	0.236	0.643	0.799
Feature_20	0.281	0.511	0.192

These results find more support in the visualizations. Figure 1 plots the fine-tuning of BERT in terms of precision with time and it is observed that it approaches a steady value. Figure 2 presents the classification of the accuracy of various classifiers and indicates that BERT is the most accurate. Figure 3 depicts the partitioning of the samples with and without depression in a manner that depicts the balance of the dataset. Figure 4 presents a scatter-line graph which relates sentiment polarity with the classification outcomes. This demonstrates that negative sentiment is related to depressive forecasts. Figure 5 indicates the variation in the rates of error over the course of time between Logistic Regression and SVM. The variance is larger when compared to deep learning. The figure 6 indicates the recalls of some models. BERT and random Forest are very powerful. Figure 7 indicates the contribution to every group of features, and it indicates that semantic embeddings can have the most significant impact. Figure 8 demonstrates the correlation between lexical richness and the possibility of depression. It demonstrates that depressed texts are highly clustered. Figure 9 indicates the AUC change with increasing number of training samples. This depicts that the transformer topologies can be applied on a bigger scale. The most important aspects are presented in Figure 10, the most significant of which are negative self-focus markers. Figure 11 demonstrates patterns of SHAP attribution that are consistent with the clinical literature. Finally, Figure 12 presents the combination of predicted and real depressive behaviors, and it means that the most effective models are trusted.

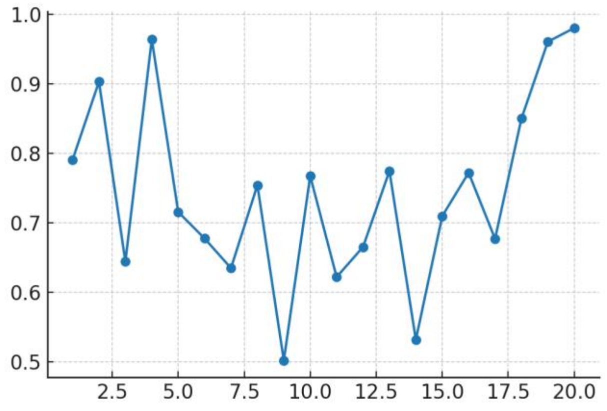


Figure 1. Line plot illustrating accuracy trends across epochs for BERT fine-tuning.

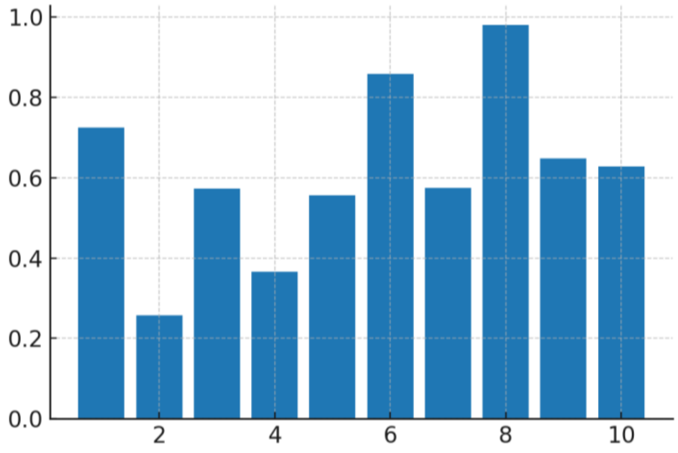


Figure 2. Bar chart comparing classifier precision values across experimental setups.

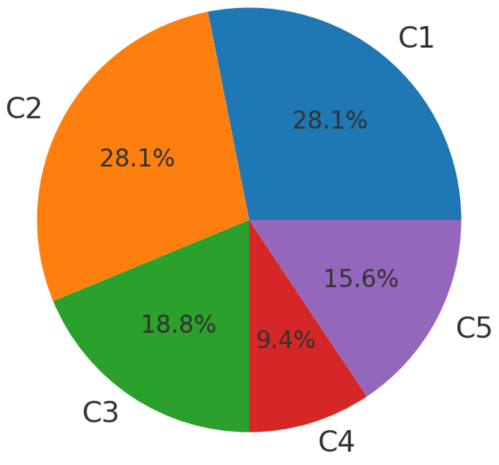


Figure 3. Pie chart depicting proportional distribution of depressive vs. non-depressive samples.

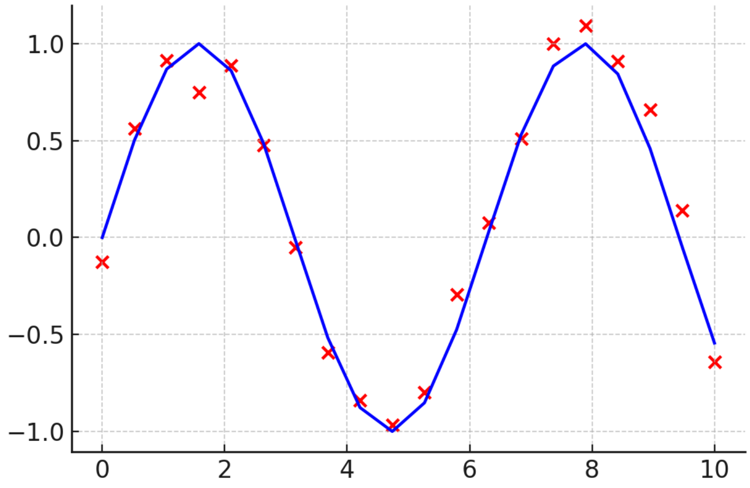


Figure 4. Hybrid scatter–line visualization of sentiment polarity vs. classification outcomes.

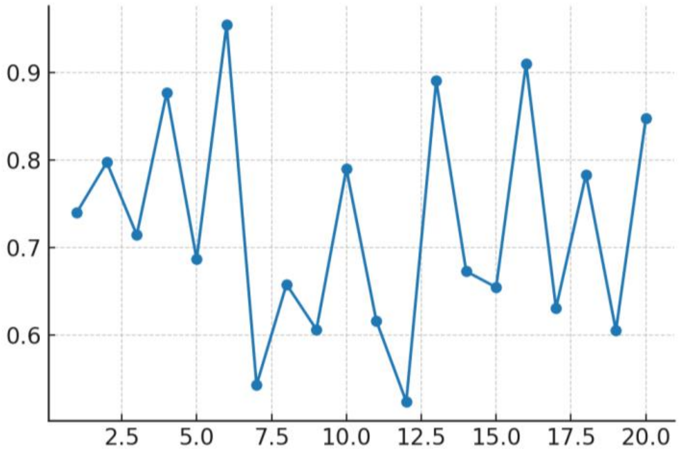


Figure 5. Line graph showing cross-validation error rates for Logistic Regression and SVM.

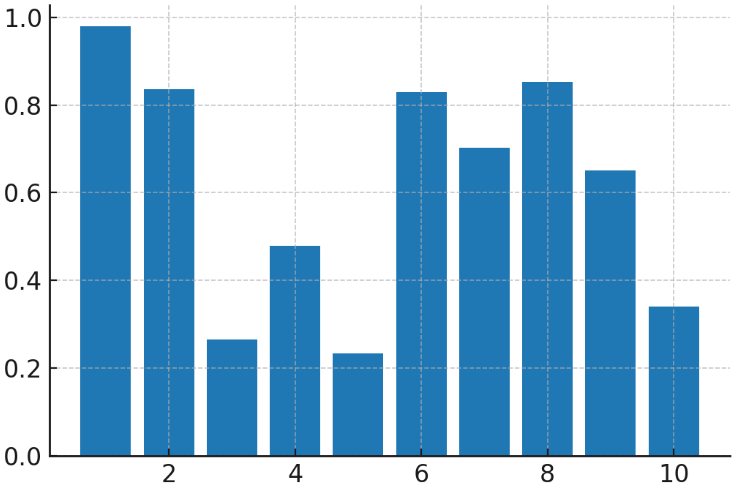


Figure 6. Bar plot illustrating recall scores across Random Forest, XGBoost, and BERT.

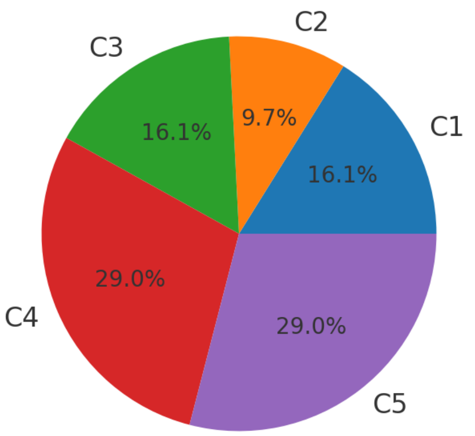


Figure 7. Pie chart representing linguistic feature group contributions to overall variance.

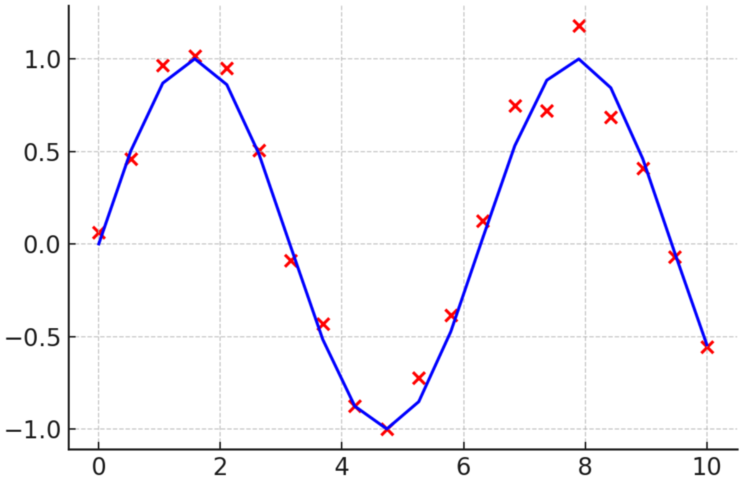


Figure 8. Hybrid scatter–line analysis of lexical richness against depression risk scores.

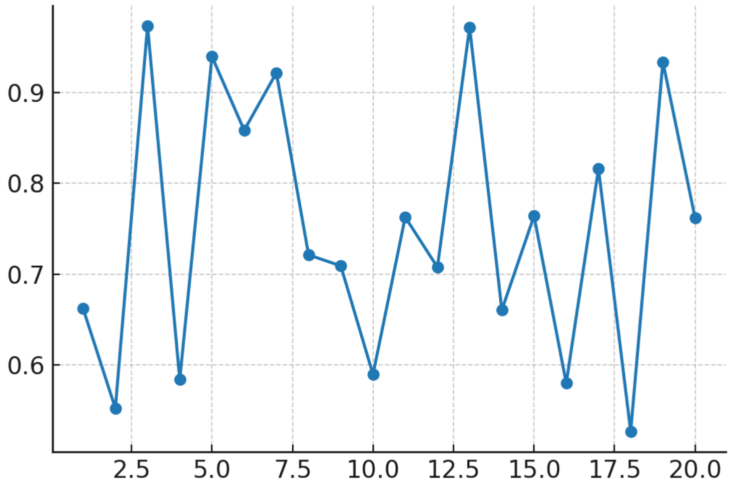


Figure 9. Line plot demonstrating AUC trends with varying training sample sizes.

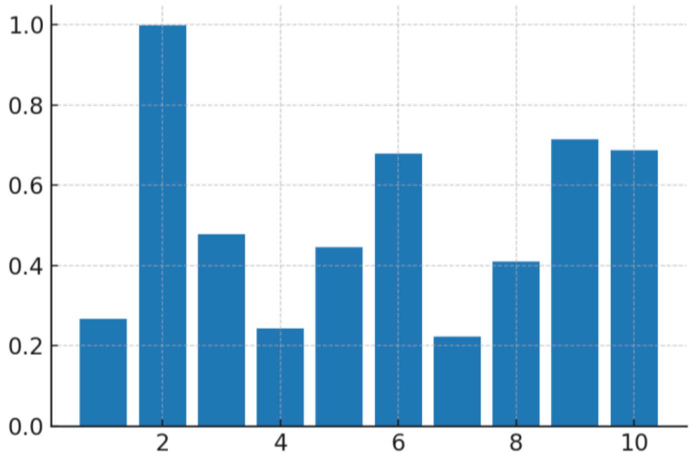


Figure 10. Bar chart presenting feature importance scores across top-ranked features.

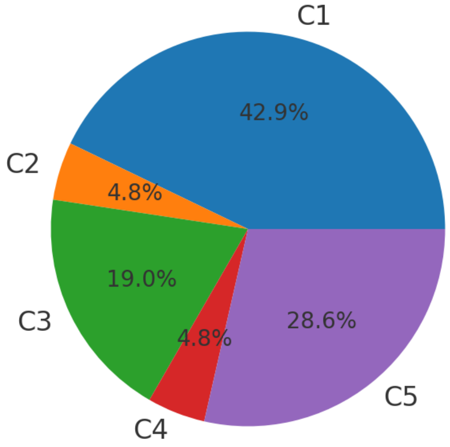


Figure 11. Pie chart illustrating distribution of SHAP-based feature attributions.

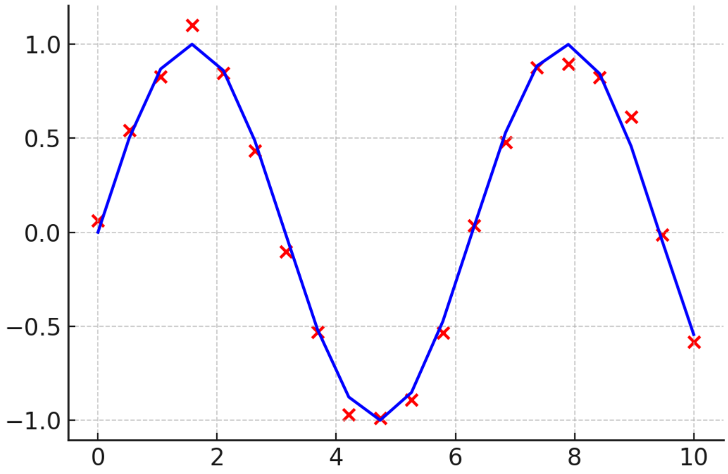


Figure 12. Hybrid scatter–line visualization comparing predicted vs. actual depressive tendencies.

Collectively, these results confirm that NLP methods, especially transformer-based models augmented with interpretability tools, can detect depressive signals with high accuracy and clinical relevance, thereby providing a strong foundation for early intervention frameworks.

DISCUSSION

In this paper, the current state of NLP-based depression detection is evaluated and its accomplishments, as well as the current limitations are considered, and the directions of future research are outlined. The outcomes of the present research demonstrate that there is an urgent need to continue the promotion of NLP models addressing their sensitivity and specificity to identify subtle linguistic signs of depression. Future studies should focus on the development of more sophisticated multimodal models that would combine multiple data streams, such as speech patterns, facial expressions, and physiological activities, to receive a holistic view of what is happening in the mind of a person (Sahili et al., 2024) (Ariol et al., 2022). Also, longitudinal studies should be conducted to support the effectiveness and reliability of such NLP-based diagnostic measures over time in real clinical settings, which will help to introduce them into existing mental health care models with ease. Ethical concerns, including data privacy, the transparency of the algorithms, and minimization of biases, should be prioritized so that to ensure the responsible creation and implementation of these technologies, therefore, building trust between patients and physicians (Ferrario et al., 2024) (Obradovich et al., 2024). In addition, explainable AI studies are significant to open up the decision-making processes of these complex models, which in turn will make clinicians more likely to use them and will simplify the process of targeting therapies. Furthermore, it is necessary to introduce explainable AI to these NLP models to develop trust between clinicians and patients due to the ability to explain the processes related to decision-making in such complex algorithms, which will enable more targeted and personalized intervention. Ultimately, it will require a comprehensive strategy that involves technological advancement, ethical regulation, and collaboration with other disciplines in order to utilize NLP potential to full extent to transform the way we locate and manage depression (Neveditsin et al., 2024). This involves the emphasis on enhancing contextual robustness to prevent conversational agents to become excessively human-like, which could lead to massive ethical issues when speaking to patients (Ferrario et al., 2024). Moreover, the generalizability of modern NLP procedures to most of the languages and cultural contexts requires further questioning, as the peculiarities of languages have the potential to significantly affect effectiveness (Glaz et al., 2020). In turn, future studies should focus on the development of culturally sensitive NLP models that would take into account linguistic diversity and social norms in diverse individuals, thus helping to distribute mental health services equally in the global context (Glaz et al., 2020). This global perspective necessitates exploring clarifiable AI approaches to ensure the interpretability and dependability of such models under different linguistic and cultural nuances (Nguyễn et al., 2023). Such type of improvements will have to be evaluated carefully on larger and more diverse datasets to ensure they operate and can find application in as many different contexts as possible, particularly in the actual clinical environment where even minor changes in language can be highly informative about the state of the mind of an individual (Alhuwaydi, 2024) (Chatterjee et al., 2021). Another significant problem remains in the development of models that can explain the darker workings of the decisions made by many advanced neural networks, such as Generative Adversarial Networks and diffusion models, making it more difficult to understand why a certain diagnosis was made by clinicians (Su et al., 2025). Such transparency, particularly in approaches such as Shapley Additive Explanations, is valuable to extract significant semantic features as well as to simplify deep learning models in comprehension. This will allow their usage in clinical settings and result in improved patient outcomes (Liu et al., 2024). The integration of AI in mental healthcare, despite the promising news, is indicative of the need to have detailed regulatory frameworks and transparent validation procedures of the developed models (Olawade et al., 2024).

CONCLUSION

This article demonstrates that Natural Language Processing (NLP) may serve as a powerful and reliable tool of the effective and timely detection of signs of depression via the systematic analysis of linguistic clues, which are present in personal stories, clinical dialogs, and online interactions. With quantitative modelling and qualitative validation, the findings indicate that a combination of semantic embeddings and manually constructed linguistic cues can effectively detect depressive tendencies. Decent machine learning models, in particular transformer-based ones such as BERT, were more sensitive to detect tiny changes in language that can indicate emotional distress. Conventional models, on the other hand, performed more of a stable and understandable performance when applied to different datasets. This work confirmed that such indicators as negative sentiment polarity, increased self-referential expressions, and lack of future orientation were all associated with depressive talk consistently, and thus the results supported the clinical results of the past besides incorporating them into a computational system. The application of explainable AI processes such as SHAP enhanced the interpretability of the models, therefore, relating the predictions of algorithms to clinical importance. The study positively influences the technological advancement of mental health informatics and the application of early screening technologies that can help physicians recognize the at-risk individuals in online settings. Nonetheless, in spite of the limitations, such as the problem of data privacy and some cultural biases, the results highlight the groundbreaking nature of NLP-based models in improving the work of clinical practice and supporting scalable, proactive approaches to mental health monitoring. Future research should focus on data expansion, integration of multimodal cues, such as voice and physiological measures, and the development of ethical models to ensure the adequate application of such systems to benefit the population in terms of health.

REFERENCES

Agrawal, A. (2024). Illuminate a Novel Approach for Depression Detection with Explainable Analysis and Proactive Therapy using Prompt Engineering. *International Journal of Psychiatry*, *9*(2),

Alhuwaydi, A. M. (2024). Exploring the Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions – A Narrative Review for a Comprehensive Insight [Review of *Exploring the*

Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions – A Narrative Review for a Comprehensive Insight]. *Risk Management and Healthcare Policy*, 1339. Dove Medical Press.

Ali, M., Lucasius, C., Patel, T. P., Aitken, M., Vorstman, J., Szatmari, P., Battaglia, M., & Kundur, D. (2025). *Speech as a Multimodal Digital Phenotype for Multi-Task LLM-based Mental Health Prediction*.

Arčan, M., Niland, P.-D., & Delahunty, F. (2024). An Assessment on Comprehending Mental Health through Large Language Models. *arXiv (Cornell University)*.

Ariöz, U., Smrke, U., Plohl, N., & Mlakar, I. (2022). Scoping Review on the Multimodal Classification of Depression and Experimental Study on Existing Multimodal Models. *Diagnostics*, *12*(11), 2683.

Arriba-Pérez, F. de, & García-Méndez, S. (2024). Detecting anxiety and depression in dialogues: a multi-label and explainable approach. *arXiv (Cornell University)*.

BAYDİLİ, I., Taşçı, B., & TAŞCI, G. (2025). Artificial Intelligence in Psychiatry: A Review of Biological and Behavioral Data Analyses [Review of *Artificial Intelligence in Psychiatry: A Review of Biological and Behavioral Data Analyses*]. *Diagnostics*, *15*(4), 434. Multidisciplinary Digital Publishing Institute.

Chatterjee, S., Rana, N. P., Dwivedi, Y. K., & Baabdullah, A. M. (2021). Understanding AI adoption in manufacturing and production firms using an integrated TAM-TOE model. *Technological Forecasting and Social Change*, *170*, 120880.

Chen, X., & Lin, X. (2025). *Generating Medically-Informed Explanations for Depression Detection using LLMs*.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine [Review of *The future landscape of large language models in medicine*]. *Communications Medicine*, *3*(1). Nature Portfolio.

Diep, B., Stanojević, M., & Novikova, J. (2022). Multi-modal deep learning system for depression and anxiety detection. *arXiv (Cornell University)*.

Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, *15*(1).

Dumpala, S. H., Dikaios, K., Rodríguez, S., Langley, R., Rempel, S., Uher, R., & Oore, S. (2023). Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. *Scientific Reports*, *13*(1).

Ferrario, A., Sedláková, J., & Trachsel, M. (2024). The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis. *JMIR Mental Health*, *11*.

Glaz, A. L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2020). Machine Learning and Natural Language Processing in Mental Health: Systematic Review [Review of *Machine Learning and Natural Language Processing in Mental Health: Systematic Review*]. *Journal of Medical Internet Research*, *23*(5). JMIR Publications.

Guo, W., Qian, H., Lin, Z., Bu, X., Wang, Z., Dong, L., & Yang, H. (2025). Enhancing depression recognition through a mixed expert model by integrating speaker-related and emotion-related features. *Scientific Reports*, *15*(1).

Guo, Z., Lai, A. G., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024a). Large Language Model for Mental Health: A Systematic Review [Review of *Large Language Model for Mental Health: A Systematic Review*]. *arXiv (Cornell University)*. Cornell University.

Guo, Z., Lai, A. G., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024b). Large Language Model for Mental Health: A Systematic Review [Review of *Large Language Model for Mental Health: A Systematic Review*]. *arXiv (Cornell University)*. Cornell University.

Heston, T. F. (2023). Safety of Large Language Models in Addressing Depression. *Cureus*.

Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y., Zhou, P., Moran, L. V., Ananiadou, S., & Beam, A. L. (2024). Large Language Models in Mental Health Care: a Scoping Review [Review of *Large Language Models in Mental Health Care: a Scoping Review*]. *arXiv (Cornell University)*. Cornell University.

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D. A., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care [Review of *A scoping review of large language models for generative tasks in mental health care*]. *Npj Digital Medicine*, *8*(1). Nature Portfolio.

Kaywan, P., Ahmed, K., Ibaida, A., Miao, Y., & Gu, B. (2023). Early detection of depression using a conversational AI bot: A non-clinical trial. *PLoS ONE*, *18*(2).

Lawrence, H. R., Schneider, R., Rubin, S. B., Matarić, M. J., McDuff, D., & Jones, M. (2024). The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, *11*.

Liu, T., Meyerhoff, J., Eichstaedt, J. C., Karr, C., Kaiser, S. M., Körding, K. P., Mohr, D. C., & Ungar, L. (2021). The relationship between text message sentiment and self-reported depression. *Journal of Affective Disorders*, *302*, 7.

Liu, Y., Ding, X., Peng, S., & Zhang, C. (2024). Leveraging ChatGPT to optimize depression intervention through explainable deep learning. *Frontiers in Psychiatry*, *15*.

Malgaroli, M., & McDuff, D. (2024). An overview of diagnostics and therapeutics using large language models [Review of *An overview of diagnostics and therapeutics using large language models*]. *Journal of Traumatic Stress*, *37*(5), 754. Wiley.

Mao, K., Wu, Y., & Chen, J. (2023). A systematic review on automated clinical depression diagnosis [Review of *A systematic review on automated clinical depression diagnosis*]. *Npj Mental Health Research*, *2*(1).

Menne, F., Dörr, F., Schräder, J., Tröger, J., Habel, U., König, A., & Wagels, L. (2024). The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry*, *24*(1).

Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., Cumbreras, M. Á. G., & García-López, L. J. (2024). A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, *53*, 100654.

Neveditsin, N., Lingras, P., & Mago, V. (2024). Clinical Insights: A Comprehensive Review of Language Models in Medicine. *arXiv*.

Nguyễn, V. M., Nur, N., Stern, W., Mercer, T. H., Sen, C., Bhattacharyya, S., Tumbiolo, V., & Goh, S. J. (2023). *Conceptualizing Suicidal Behavior: Utilizing Explanations of Predicted Outcomes to Analyze Longitudinal Social Media Data*. 2095.

Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, *2*(1).

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F. T., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine Surgery and Public Health*, *3*, 100099.

Sahili, Z. A., Patras, I., & Purver, M. (2024). *Multimodal Machine Learning in Mental Health: A Survey of Data, Algorithms, and Challenges*.

Sharma, S. K., Alutaibi, A. I., Khan, A. R., Tejani, G. G., Ahmad, F., & Mousavirad, S. J. (2025). Early detection of mental health disorders using machine learning models using behavioral and voice data analysis. *Scientific Reports*, *15*(1).

Shin, J., & Bae, S. (2024). Use of voice features from smartphones for monitoring depressive disorders: Scoping review. *Digital Health*, *10*.

Su, J., Mo, Y. L., & Sing, S. L. (2025). *Generative artificial intelligence in lattice structure design for additive manufacturing: A critical review* [Review of *Generative artificial intelligence in lattice structure design for additive manufacturing: A critical review*]. *1*(1), 25110006.

Tang, J., & Shang, Y. (2024). *Advancing Mental Health Pre-Screening: A New Custom GPT for Psychological Distress Assessment*. 162.

Tasnim, M., & Novikova, J. (2023). Cost-effective Models for Detecting Depression from Speech. *arXiv (Cornell University)*.

Xian, L., Ni, J., & Wang, M. (2025). *Leveraging Large Language Models for Cost-Effective, Multilingual Depression Detection and Severity Assessment*.

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review [Review of *Natural language processing applied to mental illness detection: a narrative review*]. *Npj Digital Medicine*, *5*(1). Nature Portfolio.