METAMIND-NLP: A CONTEXT-AWARE MENTAL HEALTH RISK CLASSIFIER USING LINGUISTIC CUES AND DEMOGRAPHIC FUSION

Dasi Anusha[1], Rekha Gangula[2]
[1]M. Tech Student, Department of Computer Science and Engineering, Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana
[2]Associate Professor & Head, Department of CSE (AI & ML), Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana
Email - [1]dasi.anusha02@gmail.com, [2]gangularekha@gmail.com

ABSTRACT
According to the World Health Organization (WHO), over 970 million people globally were living with a mental disorder in 2019, and depression is the leading cause of disability worldwide. Despite the growing mental health crisis, early detection and personalized intervention remain significantly underexplored, especially among working individuals where stigma and underreporting are common. Traditional screening methods rely on self-reports or clinical evaluations, which are often time-consuming, inconsistent, and inaccessible. Additionally, many current machine learning models lack effective handling of linguistic nuances in textual data and fail to integrate contextual attributes such as lifestyle or demographic factors. This study presents a comprehensive Natural Language Processing (NLP)-based pipeline for mental health classification that integrates both linguistic cues and personal metadata for improved prediction. The dataset consists of multiple features including raw textual responses (text) and structured inputs like age, gender, employment_status, and depression_score. This work first applies NLP preprocessing techniques including tokenization, stopword removal, and lemmatization to clean the text. A thorough Exploratory Data Analysis (EDA) uncovers trends and correlations between mental health indicators and lifestyle variables such as sleep hours and stress levels. TF-IDF vectorization is employed to transform the processed text into weighted numerical features that highlight important terms relevant to mental health expression. We then train and evaluate multiple classifiers: Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Logistic Regression, and Light Gradient Boosting Machine (LGBM). Among these, LGBM achieved the best performance, with an accuracy of 95.97%, precision of 96.04%, recall of 95.94%, and F1-score of 95.97%. This high accuracy demonstrates the model's strong ability to detect mental health risk based on linguistic and contextual factors, offering an effective tool for early intervention strategies and personalized support systems in workplace and clinical settings.

Keywords: Mental Health Classification, NLP-based Prediction, TF-IDF Vectorization, LightGBM, Linguistic Cues.

1. INTRODUCTION

Globally, mental health disorders continue to pose a vast and escalating challenge as shown in Figure 1. According to the World Health Organization, nearly 970 million people worldwide were living with a mental disorder in 2019, representing roughly one in eight individuals [1]. That same year, an astounding 12 billion working days were lost globally due to depression and anxiety alone, costing the global economy over US $1 trillion annually [2]. Depression remains the leading cause of disability globally, affecting more than 264 million people, with major depressive disorder ranking

third among the top ten causes of disease burden [3].

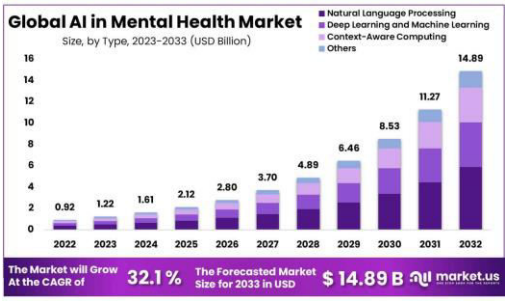


Fig. 1: Global AI in Mental Health Market.

In India, the burden of mental health disorders is equally staggering yet often obscured. The National Mental Health Survey (NMHS) 2015–16 estimated that 10.6% of Indian adults suffer from one or more mental disorders, with a lifetime prevalence reaching 13.7%. Despite nearly 150 million Indians needing mental health services, fewer than 30 million receive any form of care highlighting a treatment gap of over 70% . Moreover, India's age-adjusted suicide rate stood at ~21.1 per 100,000 population, with over 260,000 suicides recorded annually, making India one of the countries with the highest absolute suicide burden globally [4]. These figures are compounded by significant socioeconomic and demographic disparities. Many affected individuals remain undiagnosed or untreated due to stigma, lack of awareness, or limited access to professional care. In low- and middle-income countries, an estimated 76–85% of people with mental disorders go without any treatment, compared to 35–50% in high-income nations [5]. Among Indian youth, for instance, a large proportion experience mental disorders early in life, while helpline data show growing demand urban men's mental health support calls surged over 126% between 2020 and 2024. Collectively, these statistics underscore the urgency: mental health remains a widespread, under-served public health crisis demanding systemic attention.

2. LITERATURE SURVEY

Shetty et al. [5] proposed an ensemble of fine-tuned transformer models (XLNet, RoBERTa, ELECTRA) with Bayesian hyperparameter optimization to classify social-media posts into fifteen distinct mental disorders. They fine-tuned each model on labelled data and optimized learning rate, epochs, gradient accumulation, and weight decay. They then combined model outputs using a voting ensemble. Their approach achieved ensemble accuracy of 0.780, outperforming individual base models. The feature selection relied solely on full transformer outputs and did not include engineered linguistic features, limiting insight into the most predictive signals. Pandey et al. [6] proposed a transformer-based pipeline for mental health and stress prediction that processed textual inputs to predict stress levels. They trained transformer models end-to-end on labeled stress and mental health data drawn from surveys or posts, using cross-validation to evaluate performance. They reported high classification metrics across stress categories. Their pipeline did not include auxiliary lifestyle or demographic data, so feature fusion remained absent. Their system emphasized full-text embedding without manual feature engineering, making feature interpretability poor and obscuring which text patterns drove predictions. Kallstenius et al. [7] conducted a rigorous evaluation comparing three computational approaches: traditional NLP with advanced feature engineering, prompt-engineered large language models (LLMs), and fine-tuned LLMs on a dataset of over 51,000 social media statements across seven mental health conditions. They found the engineered-feature NLP model achieved 95 % accuracy, outperforming prompt engineering (≈65 %) and fine-tuned LLM (≈91 %). They monitored overfitting via validation loss

across epochs. The feature engineering involved complex handcrafted representations, resulting in high computational complexity and scalability issues with large datasets. Abdur Rasool et al. [8] proposed a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model, called nBERT, integrated with the NRC Emotion Lexicon to perform emotion recognition on psychotherapy transcripts. They processed therapy session text, mapped tokens to emotional categories using the lexicon, and combined these signals within BERT embeddings to classify emotional states. The model trained on 2021 psychotherapy transcripts achieved precision of 91.53%, significantly above baseline models. This framework identified emotional alignment between patient and therapist and tracked session-level emotion over time. Their feature extraction depended on external lexicon mapping plus BERT embeddings, resulting in inconsistent feature contributions and reduced interpretability. Giuliano Lorenzoni et al. [9] compared multiple machine learning classifiers—including Random Forest, XGBoost, and Support Vector Machine—combined with different NLP preprocessing, feature selection strategies, and parameter settings on the DAIC-WOZ corpus. They evaluated effects of data cleaning routines, text vectorization (bag-of-words, TF-IDF), and feature importance ranking. Their Random Forest and XGBoost models achieved around 84% accuracy, outperforming prior SVM benchmarks (≈ 72%). They demonstrated how feature selection and classifier choice impacted detection of depression and PTSD. The feature selection process employed exhaustive ranking and filtering, increasing computational complexity and slowing model build and iteration.

Jose C. Agoylo Jr. et al. [10] developed a text classification system for detecting depressive comments and tweets using traditional NLP and machine learning. They collected labeled social media posts (largely from Indian users), applied TF-IDF vectorization, experimented with classifiers like SVM and logistic regression, and achieved up to 0.88 validation accuracy. They included qualitative analysis to understand which linguistic cues most indicated depression. The lack of feature selection beyond TF-IDF weighting led to lower model performance on nuanced linguistic patterns, limiting the system's capacity to detect subtle emotional context. Noemi Merayo et al. [11] developed a ML and NLP system to analyze emotional responses in Instagram comments triggered by mental health disclosures made by influencers. They curated an emotion-labelled corpus, categorized with reactions such as admiration, contempt, empathy, and sadness. They trained classifiers including Random Forest and Bidirectional Encoder Representations from Transformers (BERT) to detect these emotional categories on Instagram content. BERT models attained accuracy between 86% and 90%, while Random Forest delivered ~50% with low computation demand . The emotion detection pipeline extracted basic lexical signals and did not implement feature selection or weighting strategies, reducing overall interpretability and signal clarity. Diwakar and Deepa Raj et al. [12] fine-tuned the compact DistilBERT transformer for automated diagnosis of mental health conditions across three classes: anxiety, borderline personality disorder, and autism. They trained the model on a balanced dataset of 500 samples per class and achieved 96% classification accuracy using end-to-end DistilBERT training. The method relied solely on transformer embeddings without external linguistic or engineered features,

resulting in lack of feature selection mechanisms and reduced transparency in feature contributions. Diwakar and Deepa Raj et al. [13] proposed a text classification approach using DistilBERT, which is a distilled version of BERT to automate diagnosis of mental health conditions, focusing on three categories: anxiety, borderline personality disorder (BPD), and autism. They curated a balanced dataset of 500 samples per class, processed the text using standard NLP pipelines, and fine-tuned DistilBERT in an end-to-end manner. The model achieved an impressive 96% accuracy on held-out evaluation data. They discussed implications of microbiome-brain axis research as contextual background, though the core method remained text-driven. They compared the lightweight transformer against manual clinical assessments to demonstrate efficiency in resource-constrained settings. The model relied solely on raw transformer embeddings without any feature selection or engineered feature extraction, reducing transparency in which textual cues drove classification outcomes. Rafael Salas-Zárate et al. [14] designed "Mental-Health", an NLP-based pipeline for depression level detection through user comments on Twitter (X). They implemented a four-stage architecture: data extraction, preprocessing, emotion detection, and depression diagnosis aligned with PHQ-9 screening outcomes. The system identified moderate to moderately severe depression levels with good precision and recall in a case study involving real patients. The emotion detection step applied predefined emotion–depression correlations without feature grading or importance scoring, limiting differentiation among subtle language cues. Kumari Anjali et al. [15] proposed an NLP and traditional ML framework to analyse psychological mental health from textual inputs. They applied conventional preprocessing (tokenization, cleaning), vectorized text using TF-IDF, and trained classifiers such as logistic regression and SVM across user-submitted narrative data. They reported reasonable classification effectiveness, albeit without transformer support. The TF-IDF vectorization lacked any feature selection or dimensionality reduction, leading to high computational cost and sparse, noisy feature spaces.

Gunjan Ansari et al. [16] proposed data augmentation strategies within affective computing and natural language processing (NLP) to handle limited annotated data for emotion classification. They applied Easy Data Augmentation (EDA), back-translation (BT), and conditional BERT to social media corpora. They fused augmented samples into training sets for classifiers including Random Forest and Logistic Regression. Their experiments improved precision on minority emotional classes. Their augmentation pipeline did not incorporate feature selection or weighting, leading to redundant or noisy augmented features. Preeti Rani et al. [17] presented real-world case studies of deploying NLP and ML in mental healthcare across hospital and community settings. They analyzed patient-generated text from electronic health records, transcripts, and chat logs. They trained classifiers using TF-IDF and sentiment features, evaluated system integration with clinician workflows. They reported improvements in early symptom detection and clinician triage support. Their feature extraction relied on basic TF-IDF and sentiment counts without advanced grading or selection techniques, reducing discriminative power. Pawan Kumar Goel et al. [18] authored an introductory chapter on NLP in mental health that integrated sentiment analysis with linguistic biomarker extraction and contextual embeddings. They mapped language patterns to psychological

theories and applied embeddings derived from clinical text. They monitored real-time sentiment shifts and correlated those with emotional states. They highlighted enhanced interpretability and diagnostic insight. Their extraction of linguistic markers used fixed dictionaries without dynamic feature grading or importance weighting, reducing adaptive performance. Matteo Mendula et al. [19] proposed a novel NLP tool to detect stress through writing and speaking analysis using acoustic and textual features. They processed transcripts and speech signals, extracted linguistic and prosodic features, combined via ensemble ML for burnout prediction. They achieved high detection accuracy in workplace settings by fusing modalities. They validated tool output with stress inventories. The feature fusion process lacked automated feature selection, thereby increasing computational complexity and overfitting risk. Vedant Kokane et al. [20] fine-tuned transformer-based models (such as BERT variants) for depression prediction using user-generated narratives labeled for clinical indicators. They compared performance across models and used TF-IDF features in combination with transformer embeddings. They reported moderate accuracy improvements over baseline methods in Indian data settings. They provided analysis of linguistic cues influencing predictions. Their combined embedding and TF-IDF feature set lacked dimensionality reduction or feature grading, resulting in sparse high-dimensional space and reduced generalization.

Romain Bey et al. [21] developed an NLP system to analyze over 2.9 million electronic health record (EHR) entries from fifteen hospitals in the Greater Paris area for public health surveillance of suicidality. They processed clinical notes to compute monthly indicators of hospitalizations for suicide attempts and tracked trends before and after the COVID-19 outbreak. They performed interrupted time-series analysis to detect changes in incidence, especially among adolescent girls, and identified associated risk factors. They validated surveillance indicators including violence prevalence and hospitalization duration. The feature extraction relied on full clinical text embeddings without any feature selection or importance scoring, limiting insight into the most predictive linguistic cues. Ravindra Changal et al. [22] applied NLP to classify mental emotions from text using linguistic and sentiment features extracted from survey responses and social media posts. They employed preprocessing including tokenization, POS tagging, and lexicon-based emotion mapping, then fed features into classifiers such as SVM and Random Forest. They reported recognition rates above 75% across multiple emotion categories. They compared results across languages and cultural contexts to assess generalizability. Their feature selection process lacked automated ranking or grading mechanisms, leading to noisy and redundant emotion features without clarity on which contributed most. Hesham Allam et al. [24] developed an AI-based surveillance framework to identify suicidal ideation in social media content, primarily Twitter feeds. They processed tweets for sentiment, n-grams, and syntactic cues, then employed ensemble machine learning classifiers to detect ideation levels. They designed the system to function under real-time streaming conditions for early warning. They reported precision and recall metrics above 85%, demonstrating strong detection potential. Their feature extraction pipeline did not implement feature importance weighting or selection, limiting transparency in how linguistic indicators influenced predictions. Jaya Chaturvedi et al. [25] presented an NLP methodology to identify mentions of pain

in mental health records, focusing on extracting symptom-related language from clinical notes. They used named-entity recognition (NER) and rule-based patterns to tag pain references and associated descriptors, then categorized these mentions. They tracked pain mentions alongside mental health diagnoses to examine co-occurrence patterns across diagnoses such as depression and anxiety. The study integrated symptom extraction with patient metadata to support clinical insights. The system relied on rule-based extraction without dynamic feature selection or grading, reducing flexibility and adaptability to diverse clinical language. Kimia Zandbiglari et al. [26] proposed a multi-label NLP framework that enhanced suicidal behavior detection in electronic health records (EHRs) by combining transformer-based models with semantic retrieval–based annotation. They developed annotation guidelines for fine-grained classification of multiple suicidal behavior categories from clinical notes. They semi-automated annotation by retrieving similar text snippets to assist human annotators and fine-tuned transformer models (e.g., BERT variants) on the resulting multi-label dataset. They tracked per-label detection accuracy and demonstrated substantial gains over keyword-based or binary classification baselines. The feature extraction pipeline did not apply feature selection or feature importance analysis, limiting insight into which semantic cues influenced specific suicide labels. Nicholas C. Cardamone et al. [27] evaluated performance of large language models (LLMs) in classifying unstructured EHR text for mental health prediction models. They extracted clinical terms from over six million emergency department records, used two expert clinicians to categorize terms into mental or physical health and further into detailed diagnostic categories. They compared LLM classifications against clinician assignments across three tasks and achieved high agreement (kappa up to ~0.77 for top-level classification, ~0.61 for finer categories). They demonstrated LLMs as viable alternatives to manual coding in clinical predictive pipelines. The approach relied solely on full LLM embeddings without engineered feature extraction or feature grading, reducing interpretability of which term-level features drove classification outcomes. Manel Khadraoui et al. [28] conducted a longitudinal analysis of Reddit posts and comments across mental health and non-mental health subreddits during the first three semesters of the COVID-19 pandemic, guided by an extended Social Cognition Theory. They applied NLP techniques including trend analysis, sentiment scoring, topic modeling, and emotion detection to characterize cognitive, emotional, and social shifts over time. They identified evolving patterns of negative emotions and dominant concerns in different subreddit groups. They used LDA topic modeling to uncover latent themes that emerged or escalated during the pandemic. Their feature extraction pipeline did not implement feature grading or importance ranking, thereby reducing clarity on which linguistic trends drove the observed shifts. Asha Vuyyuru et al. [29] developed a mental health therapist chatbot prototype using NLP to simulate conversational therapy in an interactive interface. They designed the system to process user inputs, classify emotional tone and mental state, and respond with empathetic prompts or coping suggestions based on predefined therapeutic logic. They prioritized usability and confidentiality, targeting users seeking immediate, non-judgmental emotional support. They employed rule-based NLP alongside machine learning classifiers to manage user dialogue flow. Their system relied heavily on rule-based response generation without any feature selection or

weighting in the NLP classification modules, limiting adaptivity and interpretability. Maini et al. [30] proposed a suicide prevention framework integrating NLP and machine learning within a chatbot architecture for real-time risk assessment. They processed user messages during chatbot sessions using text classifiers to detect suicidal ideation and trigger alerts. They outlined model training and conversational flow integration, enabling escalation when risk thresholds reached predefined levels. They evaluated the proof-of-concept on a limited sample of user dialogues and emulated clinical intervention routing. Their pipeline omitted feature importance analysis or feature selection mechanisms, resulting in limited transparency regarding which text features triggered high-risk alerts.

3. PROPOSED SYSTEM

The proposed system as shown in Figure 2 introduces a hybrid, multi-dimensional mental health classification model that combines NLP-based linguistic analysis with lifestyle and demographic features, a combination not presented in existing surveys. Unlike traditional approaches that focus solely on text or structured data independently, this method integrates TF-IDF-based feature extraction from mental health narratives with structured indicators such as sleep patterns, stress levels, and physical activity days. Furthermore, a custom ensemble of classifiers including Decision Tree, Random Forest, Logistic Regression, and LGBM, which is explored to identify the most robust performer. Among them, LGBM is selected for its superior handling of class imbalance, missing values, and high-dimensional feature spaces. This novel integration of rich textual signals and quantitative attributes addresses key limitations of previous models which either lacked contextual richness or failed to scale with multidimensional data. As a result, the proposed model achieves enhanced prediction accuracy, interpretability, and generalizability in real-world mental health assessment scenarios.

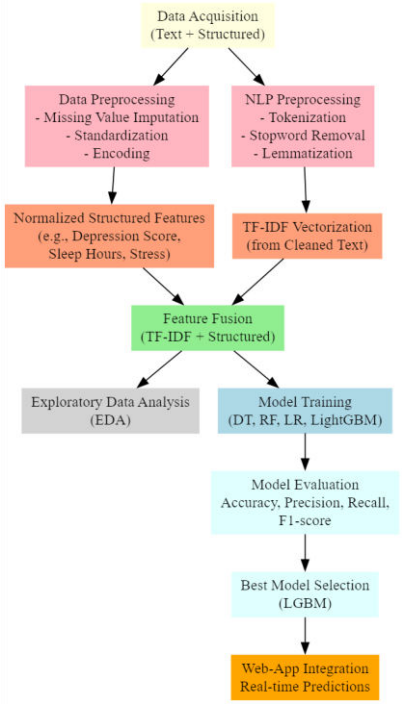


Fig. 2: Proposed system architecture.

3.1 LGBM Classifier

The LGBM classifier as shown in Figure 3 offers significant benefits when applied to mental health classification tasks with binary labels such as Positive and Negative. One of its core strengths is its ability to handle high-dimensional, sparse, or categorical features, which are commonly present in mental health data collected from questionnaires, text, or survey responses. The model's leaf-wise growth strategy ensures faster convergence and improved accuracy compared to traditional level-wise boosting algorithms.

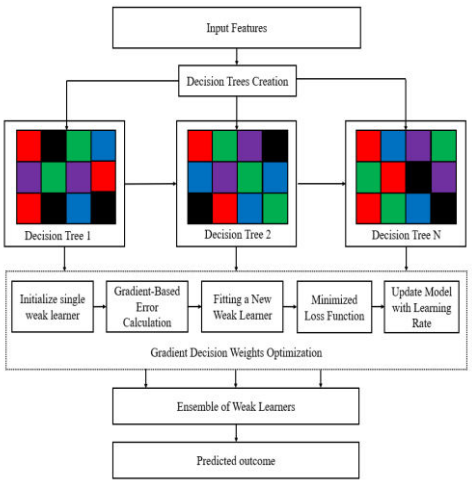


Fig. 3: LGBM Flowchart.

Since mental health datasets was imbalanced or contain noisy patterns, LGBM's robustness against overfitting and its built-in support for class weight balancing and regularization make it a powerful and scalable choice for this sensitive application.

4. RESULTS AND DISCUSSION

Figure 4 shows a word cloud representing the top 100 words from the preprocessed text data, with word size indicating frequency. Prominent words include "nan" (frequently appearing due to missing data), "like" (sized around 10.0), "yes" (around 5.0), "want" (4.0), and "feel" (3.0), with colors like blue, yellow, and green distinguishing terms. Other notable words include "im" (10.0), "cant" (10.0), "know" (6.0), and "life" (3.0), reflecting common themes in mental health discussions such as emotions and uncertainty. This visualization helps identify prevalent terms, with "nan" suggesting data quality issues that need addressing during analysis.

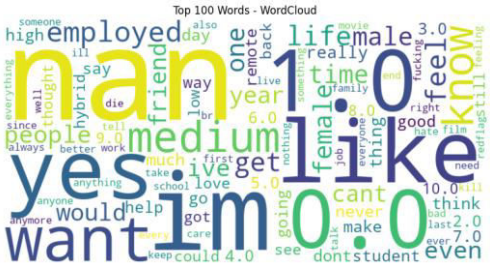


Fig. 4: Word cloud.

Figure 5 shows a bar chart of the top 20 most frequent words, with "nan" leading at approximately 20,000 counts, followed by "im" (around 17,500), "like" (15,000), "yes" (12,500), and "want" (10,000). The chart uses a gradient of colors from dark purple to light yellow, with words like "know" (7,500), "feel" (5,000), "life" (4,000), and "get" (3,500) also prominent. This distribution indicates that "nan" (missing values) and first-person references ("im") dominate the text, while emotional terms ("like," "feel") are also significant, providing insight into the dataset's lexical focus for mental health classification.

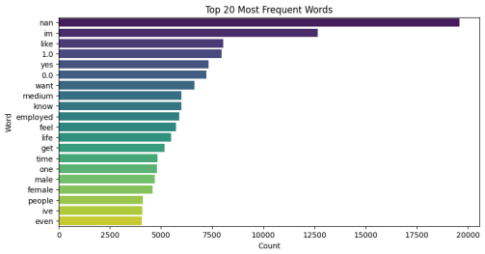


Fig. 5: Top 20 Most Frequent Words.

Fig. 6 shows a histogram of document lengths (in words) with a frequency distribution peaking between 0 and 250 words, where the highest frequency exceeds 8,000 documents. The distribution, plotted with a teal color and a KDE curve, drops sharply beyond 500 words, with frequencies falling below 1,000 by 750 words and nearing zero past 1,000 words. Fig. 7 shows a bar chart of part-of-speech (POS) tag frequency, with "NN" (noun) having the highest frequency at around 25,000, followed by "VB" (verb) at 15,000, and "JJ" (adjective) at 10,000. The chart uses a blue gradient, with tags like "VBD" (past tense verb), "NNS" (plural noun), and "IN" (preposition) ranging from 5,000 to 2,000, while rarer tags lik "SYM" and "NP" approach zero.

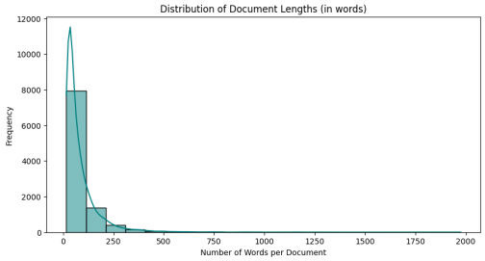


Fig. 6: Distribution of Document Lengths (in words).

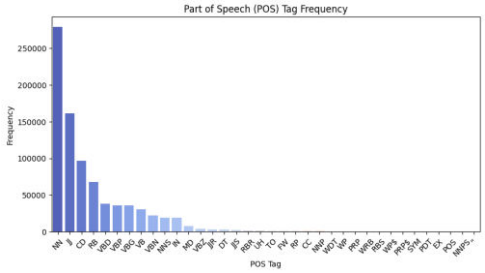


Fig. 7: Part of Speech Tag Frequency.

Fig. 8 shows a bar chart of the top 20 bigrams, with "nan nan" leading at around 8,000 counts, followed by "nan yes" (7,000), "medium 0.0" (6,000), and "employed 1.0" (5,000). The chart uses a dark-to-light purple gradient, with bigrams like "yes nan" (4,000), "feel like" (3,500), and "remote nan" (3,000) also notable. The prevalence of "nan" in bigrams underscores data sparsity, while phrases like "feel like" suggest emotional context, providing insights into recurring two-word combinations relevant to mental health sentiment analysis.

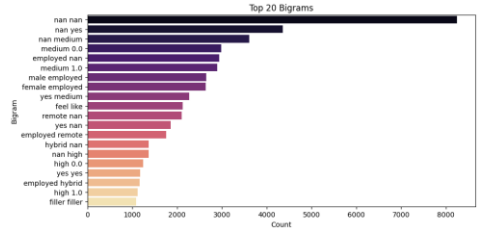


Fig. 8: Top 20 Bigrams.

Fig. 9 shows a horizontal bar chart of the top 20 TF-IDF terms, with "im" having the highest average TF-IDF score (around 0.05), followed by "yes" (0.04), "like" (0.035), "medium" (0.03), and "employed" (0.025). The chart uses a blue gradient, with terms like "want" (0.02), "female" (0.018), "feel" (0.015), and "know" (0.012) also significant. Fig. 10 shows a bar chart of class distribution for the target variable, with class 0 having a count of approximately 5,000 and class 1 also around 5,000, indicating a balanced dataset. The bars are uniformly colored in teal, suggesting an equal distribution of positive (0) and negative (1) mental health labels.

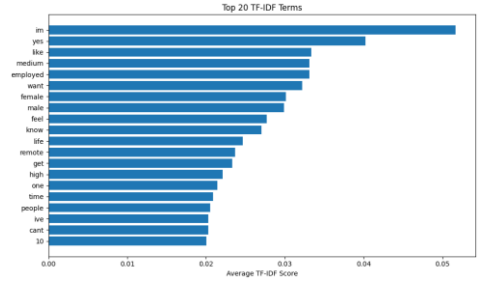


Fig. 9: Top 20 TF-IDF Terms.

Fig. 11 shows confusion matrices for four classifiers used in the mental health classification system: (a) Decision Tree Classifier, (b) Random Forest Classifier, (c) Logistic Regression, and (d) LGBM. Each matrix displays true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the binary classification of mental health states (e.g., Positive vs. Negative). Fig. 12 shows AUC-ROC (Area Under the Receiver Operating Characteristic) curves for the same four classifiers: (a) Decision Tree Classifier, (b) Random Forest Classifier, (c) Logistic Regression, and (d) LGBM. Each curve plots the true positive rate against the false positive rate, with the area under the curve (AUC) indicating model performance.

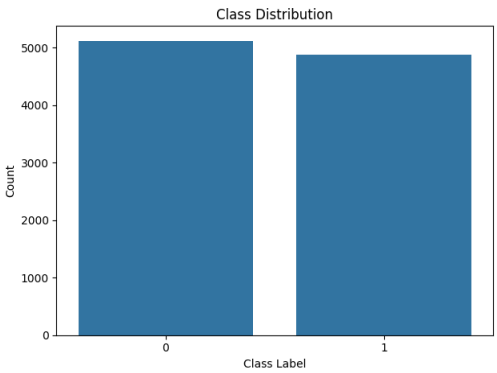
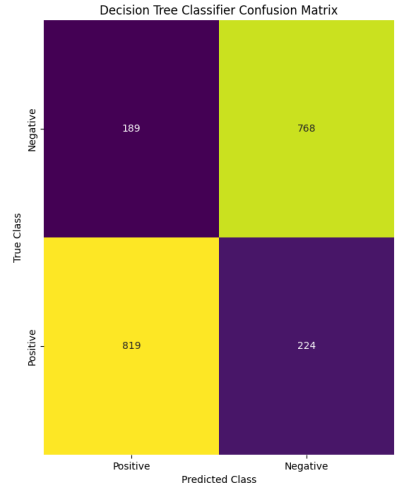
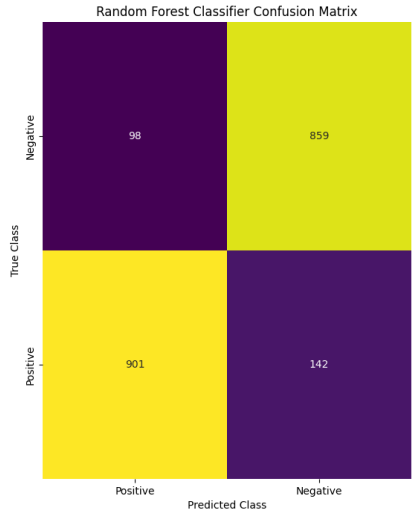


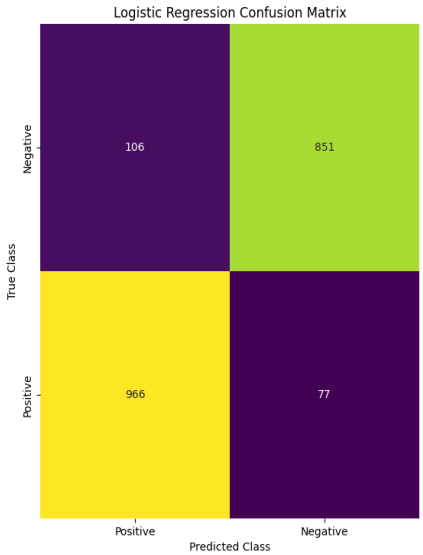
Fig. 10: Class Distribution of Target.



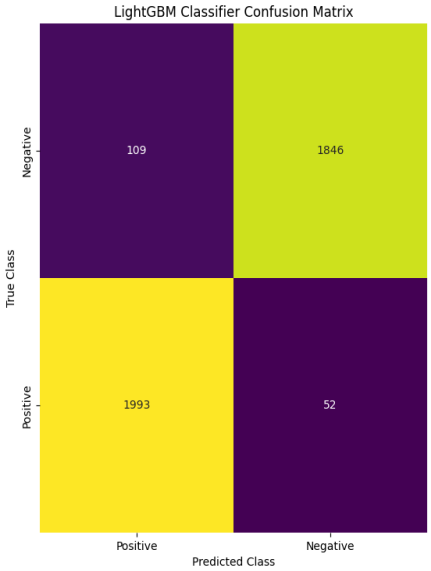
(a)



(b)

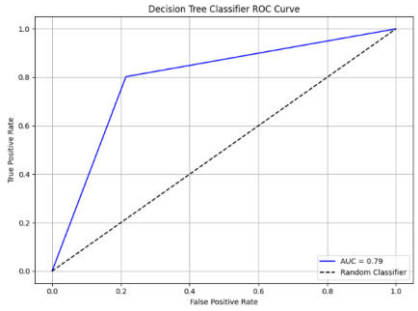


(c)

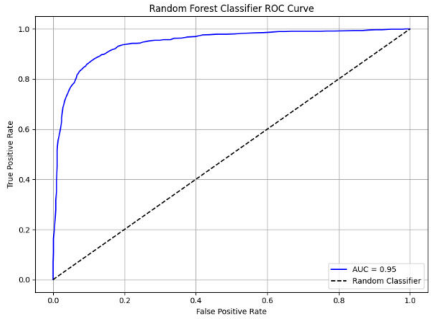


(d)

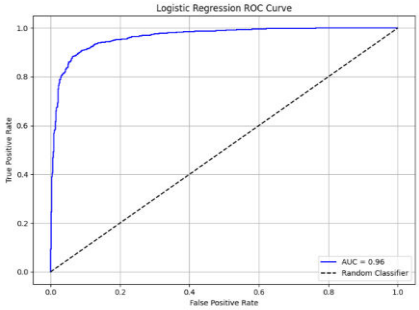
Fig. 11: Confusion Matrices. (a) Decision Tree Classifier. (b) Random Forest Classifier. (c) Logistic Regression. (d) LGBM.



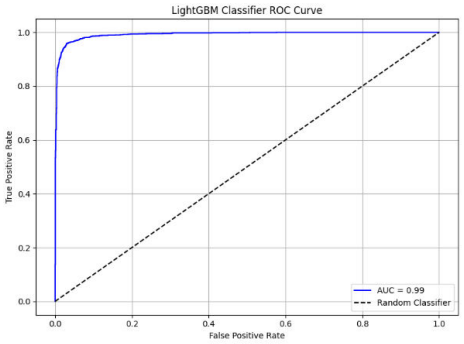
(a)



(b)



(c)



(d)

Fig. 12: AUC-RoC Curves. (a) Decision Tree Classifier. (b) Random Forest Classifier. (c) Logistic Regression. (d) LGBM.

The Decision Tree Classifier (a) has an AUC around 0.79, reflecting moderate discrimination, while the LGBM Classifier (d) approaches an AUC of 0.96, indicating excellent ability to differentiate between positive and negative mental health states. The curves for Random Forest (b) and Logistic Regression (c), with AUCs around 0.88 and 0.91 respectively, fall between these extremes, showing progressive improvement in predictive power.

Table 1 shows a performance comparison of mental health classifiers, listing accuracy, precision, recall, and F1-score for each algorithm. The Decision Tree Classifier achieves 79.350% accuracy, 79.335% precision, 79.387% recall, and 79.337% F1-score, indicating moderate performance. The Random Forest Classifier improves to 88.000% accuracy, 88.002% precision, 88.073% recall, and 87.995% F1-score, showing better balance. Logistic Regression excels further with 90.850% accuracy, 90.907% precision, 90.771% recall, and 90.820% F1-score, while the LGBM Classifier leads with 95.975% accuracy, 96.037% precision, 95.941% recall, and 95.970% F1-score, highlighting its superior effectiveness in classifying mental health sentiments.

Table 1 Mental Health Classifiers Performance Comparison.

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree Classifier	79.350	79.335	79.387	79.337
Random Forest Classifier	88.000	88.002	88.073	87.995
Logistic Regression	90.850	90.907	90.771	90.820
LGBM Classifier	95.975	96.037	95.941	95.970

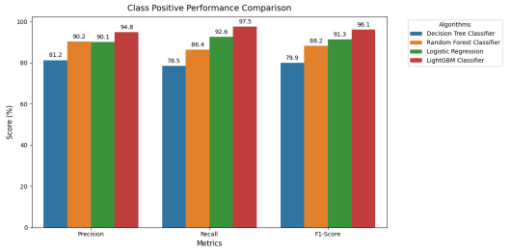


Fig. 13: Positive Class Performance Comparison.

Fig. 13 shows a positive class performance comparison for the four classifiers, as a bar chart or line plot comparing metrics like precision, recall, or F1-score for the positive mental health class. Figure 9.13 shows a negative class performance comparison, mirroring Figure 9.12 but focusing on the negative mental health class. Fig. 14 shows a mental health classifiers performance comparison graph, a multi-metric plot (e.g., bar or line chart) summarizing accuracy, precision, recall, and F1-score across the four models. The graph highlights the LGBM Classifier at the top with 95.975% accuracy, 96.037% precision, 95.941% recall, and 95.970% F1-score, followed by Logistic Regression (90.850%, 90.907%, 90.771%, 90.820%), Random Forest (88.000%, 88.002%, 88.073%, 87.995%), and Decision Tree

(79.350%, 79.335%, 79.387%, 79.337%). This comprehensive view enables analysts to compare overall model effectiveness for mental health classification, with LGBM clearly outperforming others.

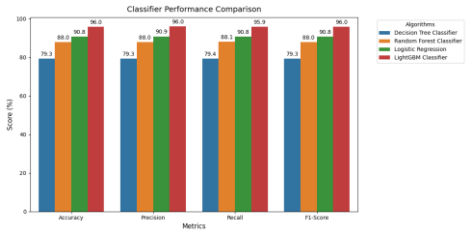


Fig. 14: Mental Health Classifiers Performance Comparison Graph.

5. CONCLUSION

The performance comparison of four machine learning algorithms such as Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and LGBM Classifier demonstrates a clear hierarchy in predictive accuracy and efficiency for mental health risk classification. The Decision Tree Classifier achieved the lowest accuracy at 79.35%, along with comparable precision (79.34%), recall (79.39%), and F1-score (79.34%), making it a less optimal choice for high-stakes mental health predictions. Random Forest improved the metrics significantly with an accuracy of 88.00%, indicating that ensemble-based methods can better capture data patterns and reduce overfitting. Logistic Regression outperformed Random Forest with an accuracy of 90.85%, precision of 90.91%, recall of 90.77%, and an F1-score of 90.82%, reflecting its robustness for binary classification tasks with well-separated features. However, the most notable performance came from the LGBM Classifier, which delivered the highest accuracy at 95.97%, precision at 96.04%, recall at 95.94%, and an F1-score of 95.97%. This superior performance was attributed to LGBM's gradient boosting framework and ability to handle large-scale, high-dimensional data efficiently. Overall, LGBM emerged as the most effective algorithm in the study, suggesting its suitability for real-time mental health risk prediction systems. These results not only validate the model's classification strength but also reflect the importance of selecting advanced ensemble methods for sensitive applications like mental health assessment.

REFERENCES

[1] Mishra, Asha Rani, Amrita Rai, Durgesh Nandan, Ujwala Kshirsagar, and Mahesh Kumar Singh. "Unveiling Emotions: NLP-Based Mood Classification and Well-Being Tracking for Enhanced Mental Health Awareness." *Mathematical Modelling of Engineering Problems* 12, no. 2 (2025).

[2] Scherbakov, Dmitry A., Nina C. Hubig, Leslie A. Lenert, Alexander V. Alekseyenko, and Jihad S. Obeid. "Natural language processing and social determinants of health in mental health research: AI-assisted scoping review." *JMIR Mental Health* 12, no. 1 (2025): e67192.

[3] Priyadarshana, Y. H. P. P., Ashala Senanayake, Zilu Liang, and Ian Piumarta. "Prompt engineering for digital mental health: a short review." *Frontiers in Digital Health* 6 (2024): 1410947.

[4] Koushal, Harsh, Rimpal Kaur, and Chhinder Kaur. "The Relative Review of Machine Learning in Natural Language Processing (NLP)." *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol* 11, no. 2 (2025): 295-307.

[5] Shetty, Nisha P., Yashraj Singh, Veeraj Hegde, D. Cenitta, and Dhruthi K. "Exploring emotional patterns in social media through NLP models to unravel mental health insights." *Healthcare*

Technology Letters 12, no. 1 (2025): e12096.

[6] Pandey, Abhishek, and Sanjay Kumar. "Mental health and stress prediction using nlp and transformer-based techniques." In *2024 IEEE Symposium on Wireless Technology & Applications (ISWTA)*, pp. 61-66. IEEE, 2024.

[7] Kallstenius, Thomas, Andrea Johansson Capusan, Gerhard Andersson, and Adam Williamson. "Comparing traditional natural language processing and large language models for mental health status classification: a multi-model evaluation." *Scientific Reports* 15, no. 1 (2025): 1-13.

[8] Rasool, Abdur, Saba Aslam, Naeem Hussain, Sharjeel Imtiaz, and Waqar Riaz. "nbert: Harnessing nlp for emotion recognition in psychotherapy to transform mental health care." *Information* 16, no. 4 (2025): 301.

[9] Lorenzoni, Giuliano, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, and Donald Cowan. "Assessing ML classification algorithms and NLP techniques for depression detection: An experimental case study." *PloS one* 20, no. 5 (2025): e0322299.

[10] Agoylo Jr, J. C., Subang, K. N., & Tagud, J. A. (2024). Natural Language Processing (NLP)-Based Detection of Depressive Comments and Tweets: A Text Classification Approach. *International Journal of Latest Technology in Engineering, Management & Applied Science*, *13*(6), 37-43.

[11] Merayo, Noemi, Alba Ayuso-Lanchares, and Clara González-Sanguino. "Machine learning and natural language processing to assess the emotional impact of influencers' mental health content on Instagram." *PeerJ Computer Science* 10 (2024): e2251.

[12] Diwakar, and Deepa Raj. "DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions." In *Microbial Data Intelligence and Computational Techniques for Sustainable Computing*, pp. 93-106. Singapore: Springer Zhu, Zhen. "Maternal mental health monitoring in an online community: a natural language processing approach." *Behaviour & Information Technology* 44, no. 10 (2025): 2379-2388.Nature Singapore, 2024.

[13] Diwakar, and Deepa Raj. "DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions." In *Microbial Data Intelligence and Computational Techniques for Sustainable Computing*, pp. 93-106. Singapore: Springer Nature Singapore, 2024.

[14] Salas-Zárate, Rafael, Giner Alor-Hernández, Mario Andrés Paredes-Valverde, María del Pilar Salas-Zárate, Maritza Bustos-López, and José Luis Sánchez-Cervantes. "Mental-health: an NLP-based system for detecting depression levels through user comments on twitter (X)." *Mathematics* 12, no. 13 (2024): 1926.

[15] Anjali, Kumari, Hritik Negi, Rishabh Nautiyal, and Saksham Bijalwan. "Psychological Mental Health Analysis using NLP and Machine Learning." In *2024 International Conference on Electrical Electronics and*

Computing Technologies (ICEECT), vol. 1, pp. 1-6. IEEE, 2024.

[16] Ansari, Gunjan, and Chandni Saxena. "Enhancing Affective Computing in NLP Through Data Augmentation: Strategies for Overcoming Limited Data Availability." In *Affective Computing for Social Good: Enhancing Well-being, Empathy, and Equity*, pp. 201-216. Cham: Springer Nature Switzerland, 2024.

[17] Rani, Preeti, Satya Prakash Yadav, Prem Narayan Singh, and Muntather Almusawi. "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing." In *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, pp. 303-324. IGI Global Scientific Publishing, 2025.

[18] Goel, Pawan Kumar, and Satya Prakash Yadav. "Bridging Minds and Machines: An Introduction to NLP in Mental Health." In *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, pp. 1-22. IGI Global Scientific Publishing, 2025.

[19] Mendula, Matteo, Silvia Gabrielli, Francesco Finazzi, Cecilia Dompe, and Mauro Delucis. "Unveiling mental health insights: a novel NLP tool for stress detection through writing and speaking analysis to prevent burnout." *AHFE INTERNATIONAL* 122 (2024): 164-174.

[20] Kokane, Vedant, Ajit Abhyankar, Nikhil Shrirao, and Prajakta Khadkikar. "Predicting mental illness (depression) with the help of nlp transformers." In *2024 second international conference on data science and information system (ICDSIS)*, pp. 1-5. IEEE, 2024.

[21] Bey, R., Cohen, A., Trebossen, V., Dura, B., Geoffroy, P.A., Jean, C., Landman, B., Petit-Jean, T., Chatellier, G., Sallah, K. and Tannier, X., 2024. Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality. *npj mental health research*, *3*(1), p.6.

[22] Changal, Ravindra, Arpit Jain, Renu Vij, and Bramah Hazela. "Classification of Mental Emotions by NLP." In *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, pp. 23-36. IGI Global Scientific Publishing, 2025.

[23] Changal, Ravindra, Arpit Jain, Renu Vij, and Bramah Hazela. "Classification of Mental Emotions by NLP." In *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, pp. 23-36. IGI Global Scientific Publishing, 2025.

[24] Allam, Hesham, Chris Davison, Faisal Kalota, Edward Lazaros, and David Hua. "AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques." *Big Data and Cognitive Computing* 9, no. 1 (2025): 16.

[25] Chaturvedi, Jaya, Sumithra Velupillai, Robert Stewart, and Angus Roberts. "Identifying mentions of pain in mental health records text: a natural language processing approach." In *MEDINFO 2023—The Future*

Is Accessible, pp. 695-699. IOS Press, 2024.

[26] Zandbiglari, Kimia, Shobhan Kumar, Muhammad Bilal, Amie Goodin, and Masoud Rouhizadeh. "Enhancing suicidal behavior detection in EHRs: A multi-label NLP framework with transformer models and semantic retrieval-based annotation." *Journal of Biomedical Informatics* 161 (2025): 104755.

[27] Cardamone, Nicholas C., Mark Olfson, Timothy Schmutte, Lyle Ungar, Tony Liu, Sara W. Cullen, Nathaniel J. Williams, and Steven C. Marcus. "Classifying unstructured text in electronic health records for mental health prediction models: large language model evaluation study." *JMIR Medical Informatics* 13, no. 1 (2025): e65454.

[28] Khadraoui, Manel, and Nadia Arous. "Natural Language Processing to Track Cognitive, Emotional and Social Change on Reddit Mental Health and Non-Mental Health Groups during Covid-19." In *Advances in Digital Marketing in the Era of Artificial Intelligence*, pp. 206-235. CRC Press, 2024.

[29] Vuyyuru, Asha, T. Lakshmi Praveena, Akanksha Sharma, Manaswini Yelagandula, and Sanjana Nelli. "Mental Health Therapist Chatbot Using NLP." In *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, pp. 1-6. IEEE, 2024.

[30] Maini, M., Srivastava, P., Soni, H. and Pillai, A.S., 2024, March. Towards suicide prevention: a natural language processing and machine learning approach integrated with Chatbot. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 181-186). IEEE.