CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

A Personality and Emotion-Aware AI Framework for Personalized Mental Health Support Using Natural Language Processing and Ensemble Learning

by

Fatima Khurshid

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2025

My work is dedicated, firstly, to the Almighty Allah, for blessing me with this opportunity, health, and ability to complete this. After Allah, this research is dedicated to my parents, my dear teachers, and my lab colleague, who encouraged and assisted me during my research.

CERTIFICATE OF APPROVAL

A Personality and Emotion-Aware AI Framework for Personalized Mental Health Support Using Natural Language Processing and Ensemble Learning

by

Fatima Khurshid

(MCS233001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Nida Adnan	NUST, Islamabad
(b)	Internal Examiner	Dr. Syed Saqib Raza Rizvi	CUST, Islamabad
(c)	Supervisor	Dr. Sabeen Masood	CUST, Islamabad

Dr. Sabeen Masood
Thesis Supervisor
October, 2025

Dr. Mohammad Masroor Ahmed
Head
Dept. of Computer Science
October, 2025

Dr. M. Abdul Qadir
Dean
Faculty of Computing
October, 2025

Author's Declaration

I, **Fatima Khurshid** hereby state that my MS thesis titled "**A Personality and Emotion-Aware AI Framework for Personalized Mental Health Support Using NLP and Ensemble Learning**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Fatima Khurshid)

Registration No: MCS233001

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Personality and Emotion-Aware AI Framework for Personalized Mental Health Support Using NLP and Ensemble Learning**" is solely my research work with no significant contribution from any other person. Small contributions/help wherever taken have been duly acknowledged, and the complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I, as an author of the above-titled thesis, declare that no portion of my thesis has been plagiarized, and any material used as a reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of the MS Degree, the University reserves the right to withdraw/revoke my MS degree, and that HEC and the University have the right to publish my name on the HEC/University website, on which names of students are placed who submitted plagiarized work.

(Fatima Khurshid)

Registration No: MCS233001

Acknowledgement

"And whoever puts all his trust in Allah, He will be enough for him." Al-Quran [65:1]. I would like to say Alhamdulillah for everything Allah has blessed me with that enabled me to reach here. I want to express my gratitude to my supervisor, Dr. Sabeen Masood, who guided me and helped me with her valuable suggestions throughout this work. And gave her valuable time, May Allah keep her in his blessings.

I am Thankful to my parents for their love, prayers, and everything that I needed, and they keep on pushing me for higher education.

Lastly, I would also like to thank everyone who has helped me along the way. Special thanks to Dr. Masroor Ahmed for increasing my knowledge and helping me with technical aspects, and for giving motivational support throughout my research journey.

(Fatima Khurshid)

Abstract

With the global rise in stress and anxiety-related disorders, the need for intelligent and scalable digital mental health support systems has become increasingly critical. Traditional approaches often lack personalization, reducing their effectiveness in addressing diverse user needs. This study presents an AI-powered framework that analyzes user-chat interactions using machine learning and natural language processing techniques to deliver tailored mental health insights. Sentence-BERT embeddings are extracted from each conversational turn and used with a Gradient Boosting Regressor to predict the Big Five personality traits, achieving a mean absolute error *MAE=0.26* and an R^2 score of *0.75*. K-Means clustering is applied to these personality vectors, followed by transformer-based detection of stress and anxiety for emotional sub-clustering. A Random Forest classifier predicts chatbot emotional responses with *97%* accuracy and macro-averaged precision, recall, and F1-scores of *0.98, 0.95, and 0.97*, respectively. To enhance response interpretability, WordNet is used to map emotional labels to behavioral synonyms. Furthermore, therapy effectiveness is predicted using multinomial logistic regression based on personality, emotional state, and message characteristics. The results demonstrate the framework's strong potential for deployment in adaptive, data-driven mental health support systems that personalize interactions, enhance engagement, and predict outcomes with high accuracy, although the study is limited by its reliance on a relatively small text-only dataset, the absence of multimodal signals such as speech or physiology, and the lack of clinical validation in real-world settings.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CBT	Cognitive Behavioural Therapy
GBR	Gradient Boosting Regressor
GMM	Gaussian Mixture Model
LM	Language Model
LSTM	Long Short-Term Memory
LWC	Linguistic Word Count
MAE	Mean Absolute Error
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
SBERT	Sentence-BERT
SMOTE	Synthetic Minority Over-sampling Technique
TF-IDF	Term Frequency–Inverse Document Frequency
T-SNE	T-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection

Symbols

c_j	Mean Cluster Embedding for Turn j
e_i	Final Embedding Vector for Sentence/Document i
p_t	Softmax-Based Emotion Prediction from Transformer Encoding
$\mathrm{RF}(\mathbf{x})$	Random Forest Prediction via Majority Voting of Decision Trees
R^2	R-Squared Coefficient
$\bar{s}$	Overall Average Silhouette Score for Clustering Evaluation
s_u	Binary Emotional State Vector for User u
$\mathrm{sil}(i)/s(i)$	Silhouette Score Analysis
$\bar{s}(k')$	Cluster-wise Silhouette Evaluation for k'
S	Silhouette Coefficient Formula for a Single Sample
$\bar{\mathbf{s}}_{C_{k'}}$	Centroid of Stress–Anxiety Vectors for Subcluster $C_{k'}$
s_t	Binary Stress–Anxiety State Vector at Time t
x_i	Combined Embedding of Context and Response
$\hat{y}_i$	Multi-Output Regression for Big Five Personality Traits
Y	Output Matrix of Big Five Personality Predictions
$\hat{y}_{\text{emotion}}$	Emotion Classification Output Label
$\hat{y}_{\text{behavior}}$	Emotion-to-Behavior Mapping via Random Synonym Selection
$\hat{y}$	End-to-End Prediction Pipeline with SMOTE and Random Forest
$\bar{\mathbf{y}}$	Mean Feature Vector Across All Samples
z	Z-Score Normalization of the Silhouette Coefficient
z_t	Integrated Personality–Stress–Anxiety Vector at time t
μ_k	Centroid Calculation for Cluster C_k in Personality Space

Chapter 1

Introduction

1.1 Background

Mental health challenges, particularly stress and anxiety, are increasing worldwide. According to estimates from the World Health Organization (WHO), more than *300* million people suffer from anxiety disorders, while millions more live in chronic stress conditions every day [1]. Despite increasing awareness, many people still face barriers in accessing timely and effective care due to stigma, high costs, and limited availability of qualified professionals [2].

These factors highlight the urgent need for accessible, affordable, and scalable support systems. From a technical perspective, Artificial Intelligence (AI) offers promising possibilities, particularly through Natural Language Processing (NLP)-driven conversational agents (chatbots).

Such systems have demonstrated encouraging results in mental health contexts by simulating empathetic dialogue and guiding users with coping strategies [3]. Current AI-based solutions provide an alternative pathway for those unable to access traditional healthcare. In this research, the term *therapy* is used in a contextual sense, referring to structured interventions such as Cognitive Behavioral Therapy (CBT) that are delivered by trained professionals.

While prior research has explored chatbot-assisted therapy delivery, the present work does not attempt to replace or automate clinical therapy. Instead, therapy is mentioned only to situate this research within the broader landscape of AI applications in mental health, while the actual focus remains on personalization, emotional awareness, and predictive modeling within conversational data.

However, existing approaches remain largely generic in nature. While they may provide supportive responses, they rarely adapt to the unique personality traits or emotional states of users. This lack of personalization reduces their effectiveness and limits sustained engagement. In addition, most models are designed as isolated modules, focusing on personality inference, stress/anxiety detection, or basic response generation, without integrating these components into a unified and adaptive framework.

To address these gaps, the present research sets out the following objectives: first, to analyze chatbot interaction data using NLP techniques to identify emotional patterns and cluster users based on inferred Big Five personality traits. This enables grouping users into psychologically meaningful clusters, forming the basis for more adaptive interaction. Second, to assess how personality traits influence the effectiveness of AI-driven responses by applying regression-based models. Together, these objectives aim to move beyond static, one-size-fits-all systems and toward psychologically aware frameworks capable of enhancing personalization, adaptability, and user trust in AI-driven mental health support.

1.2 Motivation

Although AI-powered mental health chatbots have shown potential, several critical limitations persist. First, current systems tend to deliver one-size-fits-all responses that fail to adapt to individual differences in personality, communication style, or emotional variation over time. Second, the absence of mechanisms to integrate real-time personality inference and emotion detection limits the depth of personalization that such systems can achieve. Finally, while some models incorporate

structured therapies such as CBT, their inability to dynamically adjust to a user's evolving stress or anxiety levels reduces their therapeutic value in the long run.

These gaps form the core motivation for this research. By addressing the limitations of existing work, the aim is to design an AI-driven framework capable of analyzing user interactions and providing personalized insights. Rather than offering therapeutic interventions, the system focuses on recognizing stress and anxiety patterns and adapting its analysis to the user's personality traits and emotional context.

This approach enhances the relevance, interpretability, and usefulness of the system while supporting, rather than replacing, professional mental health care.

1.3 Problem Statement

Existing AI-powered chatbots in the mental health domain often provide generic and predefined responses, overlooking the individuality of users. In the context of stress and anxiety, these systems rarely incorporate mechanisms for dynamic emotional pattern analysis or personality-based modeling. As a result, interactions remain superficial, reducing user engagement and limiting the system's ability to adapt meaningfully to diverse users.

This lack of personalization often leads to emotionally detached or robotic conversations, which can discourage users from seeking continued support. Moreover, most existing systems operate in isolated modules, failing to integrate personality and emotion analysis into a unified pipeline that can adjust responses in real time.

The problem addressed in this research is the absence of an AI-driven framework that can go beyond static, one-size-fits-all responses. There is a clear need for a system capable of identifying emotional discomfort, inferring personality traits, and forecasting contextually appropriate responses. Unlike therapeutic tools, the focus here is not on delivering clinical interventions but on creating psychologically aware interactions that enhance relevance, adaptability, and trust in AI-driven

mental health support systems. By bridging the gap between emotional understanding and personality-aware adaptation, this work aims to lay the groundwork for more human-like and responsive digital mental health support.

1.4 Research Questions

This thesis investigates the following key questions:

RQ-1: How can Natural Language Processing (NLP) be used to analyze emotional patterns in AI-driven therapy interactions?

RQ-2: How can clustering algorithms be applied to group users based on personality traits and therapy preferences?

- Can personality and stress/anxiety-based user clusters predict the sentiment or emotional tone of chatbot responses?
- How do chatbot responses' sentiments or emotions vary across different personality and stress/anxiety user clusters?
- To what extent do user personality and stress/anxiety clusters influence the behavior of AI chatbot responses?

RQ-3: What is the impact of personality traits on the effectiveness and engagement of AI-based stress and anxiety therapy?

1.5 Technological Foundations

1.5.1 Natural Language Processing

Natural Language Processing (NLP) enables the system to interpret and analyze user-generated text in a psychologically meaningful way. By applying advanced

language models such as Sentence-BERT, the system generates sentence-level embeddings that preserve semantic and emotional context. These embeddings form the foundation for downstream tasks including personality trait inference, stress/anxiety detection, and clustering, making NLP a critical enabler of this research [4].

Applicability: In this work, NLP is applied to short, real-world conversational data such as chatbot interactions. Its scope extends to extracting semantic and emotional signals that can later be modeled by ML algorithms for psychological profiling and adaptive support.

1.5.2 Deep Learning

Deep learning, particularly through transformer-based encoders, allows the system to capture contextual information from user text beyond what traditional approaches such as TF-IDF or Bag-of-Words can achieve. In this thesis, deep learning models are used to produce high-quality text embeddings that preserve both meaning and subtle emotional cues, which are vital for accurate psychological interpretation [4].

Applicability: Deep learning is applied primarily for representation learning, enabling the system to derive embeddings that serve as the input to multiple ML modules. This ensures that emotional tone, stress signals, and personality-related patterns are preserved throughout the analysis pipeline.

1.5.3 Machine Learning

Machine Learning (ML) is the central predictive and analytical engine across the entire research pipeline. ML models are applied at nearly every stage:

- Multi-output regression for Big Five personality trait inference.

- Clustering algorithms (e.g., K-Means) for grouping users into personality-based and emotion-informed subclusters and for stress and anxiety detection.
- Random Forest models for predicting chatbot responses and WordNet used for mapping emotions to behavioral categories.
- Random Forest for evaluating therapy effectiveness based on intervention outcomes.

Applicability: The extent of ML in this research is broad and foundational. Rather than serving a single role, ML integrates the entire pipeline from personality profiling to emotion-aware interaction and outcome prediction. Its applicability spans mental health support chatbots, adaptive stress/anxiety monitoring systems, and broader domains in affective computing and human–AI interaction. This integration demonstrates the potential of ML to enable scalable, personalized, and context-sensitive user experiences across diverse psychological support settings.

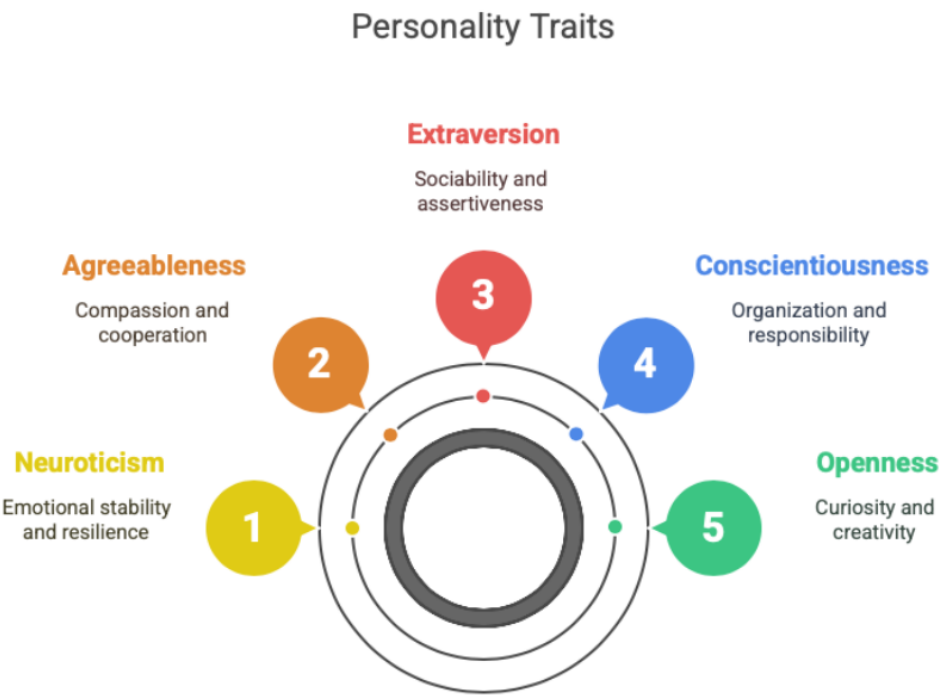

FIGURE 1.1: Personality Traits

1.6 Scope of Study

This study focuses on the design and evaluation of an AI-based support system for stress and anxiety management that integrates natural language processing, personality trait inference, and emotion-aware response prediction. The system uses sentence embeddings generated through Sentence-BERT, applies gradient boosting regression to infer Big Five personality traits, clusters users based on psychological and emotional features, and uses a Random Forest model to predict chatbot responses. The framework is further evaluated using multinomial logistic regression to understand how personality-emotion clusters influence response effectiveness. The study is limited to textual data and simulation-based chatbot interactions.

1.7 Significance of the Study

This research contributes to the growing field of intelligent digital therapy by demonstrating how psychological personalization can improve AI-driven mental health interventions. By modeling user personality traits and emotional patterns, particularly stress and anxiety, the proposed system offers a deeper, more human-centered approach to conversational therapy. It also introduces a hybrid AI pipeline that combines machine learning, deep learning, and behavioral clustering to tailor chatbot responses, thereby enhancing user engagement and therapeutic outcomes. The findings may inform future developments in emotionally intelligent virtual agents and adaptive mental health technologies.

1.8 Thesis Organization

The structure of this thesis outlines the organization of the document as follows: It provides a visual overview of the major chapters and their logical flow, showing how each component builds upon the previous one. This roadmap helps the reader understand the progression from background and literature review to methodology,

experimental results, and concluding discussions, ensuring clarity and coherence throughout the document.

1.8.1 Chapter 1 – Introduction

This first chapter provides an overview of the background, purpose, problem statement, research question, scope, significance, objectives achieved, and thesis organization of the study.

1.8.2 Chapter 2 – Literature Review

Surveys foundational and contemporary research on AI in mental health, emotional analysis, personality modeling, and chatbot design.

1.8.3 Chapter 3 – Methodology

Describes the dataset, models, embedding techniques, clustering, prediction, and regression components in detail.

1.8.4 Chapter 4 – Results and Evaluation

Presents the empirical findings, including visualizations, cluster interpretations, and regression analysis.

1.8.5 Chapter 5 – Conclusion and Future Work

Summarizes contributions, outlines limitations, and suggests directions for future studies.

Chapter 2

Literature Review

2.1 Global Mental Health Crisis and the Role of Artificial Intelligence

The World Health Organization 2023 fact sheet outlines the global impact of mental disorders, highlighting that more than 970 million people suffer from conditions such as anxiety and depression. The paper emphasizes that the mental health gap, especially in low-resource regions, demands scalable and accessible solutions. This global perspective provides the foundational rationale for exploring AI-driven tools for mental health support [5]. Yang Ni and Fanli Jia [6] analyze more than 30 mental health chatbot systems and explore their effectiveness in providing psychological support. Digital chatbots may reduce user suffering and improve mental health, according to their results. However, they also highlight the need for stronger ethical frameworks, particularly in terms of data protection, personalization, and user trust in AI-driven therapeutic systems. They emphasize that without addressing these concerns, large-scale adoption of such systems could face resistance from both users and healthcare providers. Furthermore, they suggest that future chatbot designs must integrate transparency and explainability features to build long-term user confidence.

2.2 NLP and Sentence Representation in Mental Health Systems

Reimers and Gurevych [7] introduce Sentence-BERT, an Asian network-based adaptation of the BERT architecture that produces semantically rich sentence embeddings. This model is very significant for AI-powered treatment systems, as it performs exceptionally well in semantic textual similarity tests and is frequently employed in applications such as emotion categorization, personality inference, and conversational modeling.

Marco et al. [8] present TweetEval, a thorough standard for assessing the categorization of sentiment and emotions across Twitter datasets. In emotion detection tasks, it is shown that fine-tuned transformer models perform noticeably better than conventional models, providing strong performance for multi-label classification settings that are crucial in mental health contexts.

2.3 Emotion-Aware and Affective Chatbots

Emma L. van der Schyff et al. [9] provide an emotion-aware conversation system that combines the creation of therapeutic chatbot responses with real-time emotion classification. Their technology demonstrated a 13% increase in user-perceived emotion and session satisfaction when tested in healthcare contexts, highlighting the importance of emotional computing in enhancing chatbot interactions. Rathnayaka et al. [10] provide a review of the self-directed AI-based chatbot for mental health called Bunji. Bunji guides people through everyday therapeutic conversations based on Cognitive Behavioural Therapy (CBT) using Natural Language Processing (NLP). The system's potential as a stand-alone digital support agent was confirmed when it was shown to reduce anxiety symptoms and received favourable user ratings in a field study with 112 participants. Patel et al. [11] used an intelligent chatbot for mental health to conduct a randomised controlled experiment with college students. According to PHQ-8 scores, participants'

feelings of anxiety and depression significantly improved throughout a two-week intervention. The chatbot's use of mood monitoring and Cognitive Behavioural Therapy (CBT) techniques shows that AI-driven treatments can be successful in providing temporary mental health support.

Lee et al. [12] present a deep learning–based counseling chatbot that recognises intent and classifies emotions using LSTM networks. It features a retrieval-based response engine trained on therapy-like dialogues. A trial deployment with 54 users demonstrated high engagement and a noticeable improvement in perceived therapeutic connection, underscoring the potential of emotion-sensitive chatbots. The possibility of employing AI-powered chatbots for temporary mental health assistance has been shown by a number of current systems. For example, in real-world research, self-directed tools based on Cognitive Behavioural Therapy have demonstrated potential in reducing anxiety symptoms and obtaining favourable user feedback. Similarly, over short intervention periods, mood-monitoring bots designed for college students have demonstrated quantifiable decreases in anxiety and sadness. Deep learning-based chatbots have also investigated intent classification and emotion identification to promote more engaging and therapeutic dialogues. However, despite these advances, the majority of current techniques remain narrowly focused; they largely rely on predetermined therapy scripts, short-term outcomes, or the detection of discrete emotional states. This often limits the scalability and long-term adaptability of such systems, as they fail to capture the evolving nature of human psychological well-being. Moreover, existing frameworks typically lack integration of multiple psychological dimensions such as stress trajectories, anxiety fluctuation patterns, and individual personality traits, which play a central role in shaping human emotional experiences. These limitations highlight an important research gap. The combination of long-term personality assessment with real-time emotional understanding to design highly personalised and psychologically adaptive chatbots is still underexplored. Addressing this gap requires moving beyond static models and incorporating adaptive learning strategies that account for temporal variations in user emotions and behaviours. By creating a framework that dynamically customises chatbot responses based on

individual personality traits, stress patterns, and current emotional states, this study seeks to expand the breadth, continuity, and effectiveness of digital mental health support, offering a step towards more sustainable, user-centric interventions. In doing so, it advances the field from reactive support toward proactive, personalised mental health care.

2.4 Effectiveness and Engagement with Chatbots

Li et al. [13] conducted a meta-analysis of over 30 studies examining AI-based mental health chatbots. The analysis revealed moderate effect sizes for chatbot-driven reductions in anxiety $g \approx 0.38$ and depression $g \approx 0.38$. Although effective overall, the authors noted variability in engagement and the need for further personalization in chatbot behavior.

Casu et al. [14] performed a scoping review of 64 chatbot systems in mental healthcare. They found that only a small fraction incorporated user personality traits or emotional profiles into response selection. The study recommended integrating personality-aware modeling and emotional subtext detection to improve personalization and reduce the "robotic" tone of chatbot responses.

Sim and Choo [15] investigate privacy and ethical concerns in AI-powered mental health applications. The study reviews over 40 digital therapy systems and finds that privacy-preserving technologies like differential privacy and federated learning are rarely implemented despite growing user anxiety about data misuse. The authors propose a privacy-first design framework for mental health chatbots and emphasize the need for transparent consent mechanisms.

Together, these studies highlight that while chatbot interventions can deliver measurable therapeutic benefits, their long-term success depends on balancing **clinical effectiveness, personalization, and ethical safeguards**. Current systems are effective at addressing immediate symptoms but often lack the adaptive depth required to maintain sustained user trust and engagement. Addressing these gaps

requires moving beyond short-term outcomes toward solutions that are personalized, ethically grounded, and capable of long-term therapeutic alignment.

The increasing relevance of AI-based chatbots for mental health in lowering anxiety and depressive symptoms has been confirmed by recent evaluations and assessments. Although these systems result in moderate increases in user well-being, meta-analytical research indicates that their effectiveness frequently changes depending on contextual alignment and user participation. Only a small percentage of chatbots include user personality features or emotional depth in their response tactics; the majority still depend on basic dialogue frameworks, according to broader studies of implementations. Users may perceive chatbot conversations as robotic or emotionally cold as a result of this lack of personalisation. Furthermore, many mental health platforms' designs still fall short in addressing important ethical aspects like data protection and permission procedures, which raises questions regarding long-term user acceptance and trust. The current study offers a more comprehensive and human-centered solution to these drawbacks by combining psychological profiling, emotional subtext recognition, and privacy-conscious design principles to provide chatbot interactions that are emotionally intelligent, secure, and adaptive for each user.

2.5 AI-Based Personality Trait Inference

Habib et al. [16] describe a Bi-LSTM-inspired deep learning framework based on small and medium transformer models that was trained on Reddit postings to infer users' Big Five personality traits. Their method outperformed traditional machine learning techniques such as ridge regression and significantly reduced the Mean Absolute Error (MAE) to around 0.08. The model's performance demonstrates the effectiveness of contextual semantic learning in predicting personality traits from user-generated information.

Habib et al. [17] used a Gradient Boosting Regressor on Sentence-BERT embeddings to predict the Big Five personality traits from Reddit comments. Their

model demonstrated low resource consumption, high interpretability, and achieved a Mean Absolute Error (MAE) of approximately 0.10. The study supports the use of ensemble tree-based regressors for real-time personality inference in chatbot contexts. Kunte and Panicker [18] proposed a machine learning–based approach to infer personality qualities from textual data using social media and chatbot-like logs.

Their ensemble of models, including Random Forest, Support Vector Regression, and Gradient Boosting, outperformed individual models by about 4% in terms of Mean Absolute Error (MAE), demonstrating the effectiveness of ensemble learning for predicting Big Five personality traits.

Christian et al. [19] used pre-trained Sentence-BERT embeddings with model averaging to build a multi-output regression framework that predicts all five Big Five personality traits simultaneously from text collected across multiple social media platforms. Their architecture improves training efficiency and supports real-time personality inference, making it well-suited for integration into live AI-chatbot systems for therapeutic interaction.

Zim et al. [20] proposed a hybrid deep learning model that combines transformer-based multi-target regression and transfer learning from general-domain to personality specific datasets. When evaluated on the Essays and MyPersonality datasets, their model achieved a Mean Absolute Error (MAE) below 0.09. These findings support the effectiveness of transformer-based models for accurately predicting personality traits in mental health contexts.

Bhandari et al. [21] contrasted conversation-level discourse representations with sentence-level embedding models for predicting Big Five personality traits. Their results showed that embeddings derived from entire user conversations outperformed sentence-level approaches, raising the coefficient of determination R^2 from 0.32 to 0.48. This highlights the importance of conversational context in capturing persistent behavioural patterns.

Numerous techniques have been investigated for identifying the Big Five personality characteristics from user-generated text, demonstrating notable improvements

in model precision and creative architecture. Prediction error has been considerably decreased by deep learning frameworks like transformer-based models and Bi-LSTM, which have shown exceptional performance in capturing contextual semantics. By combining both textual and network-level user information, hybrid architectures that use ensemble learning or graph-based structures have further enhanced inference. Furthermore, research has started to highlight the value of real-time inference capabilities and multi-output regression models, especially for use in live chatbot contexts. Even with these developments, there are still a number of restrictions. In conversational systems, a lot of models give performance measurements precedence over useful deployability. Furthermore, only a small number of methods take into account emotional context, real-time restrictions, or the integration of personality inference with tasks that come after, such as mental health support. The majority of models are developed and assessed independently, without reference to dynamic conversational data or therapeutic goals. In order to close this gap, the current study integrates an accurate and efficient personality inference pipeline into an AI chatbot architecture in addition to employing ensemble regression and sentence-level embeddings. The psychological relevance and therapeutic potential of digital mental health interactions are increased by this integration, which makes it possible to provide context-aware, personalised replies that are suited to both emotional states and persistent personality features.

2.6 Emotional Pattern Detection from Text

Chen and Lee [22] proposed deep learning–based models, including hybrid CNN architectures, to classify stress levels in university students using physiological signals collected during Sudoku tasks. Their models achieved an F1-score of up to 0.93, demonstrating that combining convolutional layers for feature extraction with deep contextual representations is highly effective. This work has direct implications for how therapeutic chatbots can classify emotional states from brief signals or texts. Gaballah et al. [23] conducted a comparative study on workplace stress detection, evaluating context-aware speech-based deep learning

models against traditional approaches. Using Bi-LSTM classifiers on real-world data from hospital workers, their context-aware system outperformed traditional lexicon-based methods by over 10% in F1-score, supporting the shift toward deep learning models such as BERT for real-time emotion monitoring in organizational settings. Acheampong et al. [24] used domain-specific fine-tuning of BERT on emotion-labeled text to classify emotional tones such as anger, sadness, and joy. Their transfer learning approach improved the macro F1-score by about 5% compared to generic pre-trained models. This demonstrates that adapting BERT with therapy-related or domain-specific data can significantly enhance emotion detection performance in AI-driven therapeutic systems. Ahanin et al. [25] proposed an emotion detection framework that integrates transformer-based BERT embeddings with lexicon-derived emotional features such as sentiment polarity, emoji, and hashtag cues. Their hybrid approach achieved up to 68.40% Jaccard accuracy on the SemEval-2018 dataset and improved recall for high-stress messages by 7%, demonstrating the benefit of combining learned features with rule-based emotional dictionaries for reliable mental health detection.

By applying transformer-based architectures, hybrid models, and domain-adapted embeddings, recent research in stress and emotion detection has advanced significantly. While studies concentrating on workplace and educational settings have revealed that transformer models significantly outperform traditional lexicon-based techniques, approaches that combine convolutional layers with contextual embeddings have shown high performance in detecting stress signals in short texts, such as diary entries. The capacity to recognise emotional states and high-risk stress indicators in a variety of text forms has been significantly improved by the combination of semantic elements with topic modelling and statistical metrics. Additionally, domain-specific fine-tuning has shown promise, particularly in therapeutic settings when general-purpose models are inadequate. Recent studies also highlight the importance of incorporating longitudinal behavioural trends to capture evolving emotional patterns, further enhancing predictive reliability. Despite these advancements, the majority of current systems are still concentrated on restricted, domain-specific use cases or independent classification performance.

They frequently lack the capacity to adapt in real time, generalise to different kinds of users, or incorporate emotional insights into more complex interactive systems like chatbots.

To further support a deeper understanding of behaviour, there is also little focus on integrating psychological variables like personality characteristics with emotional detection. The current work uses a multi-layered emotional analysis pipeline that includes sub-clustering, transformer embeddings, and stress/anxiety categorisation in order to close these gaps. In addition to increasing detection accuracy, this approach makes it possible to provide psychologically relevant replies inside an adaptive chatbot framework, providing more purposeful and supportive interactions in mental health settings. By embedding personality-aware insights into emotional modelling, the system ensures that user responses are not only contextually accurate but also tailored to individual differences. This alignment of psychological and emotional dimensions enhances the overall effectiveness and trustworthiness of AI-driven therapeutic support.

2.7 Clustering for Personalization

Hornstein et al. [26] conducted a systematic review of 138 studies on digital mental health interventions targeting depressive symptoms, analyzing 94 distinct interventions. They found that 66% incorporated personalization mechanisms, and personality-based content tailoring was associated with improved adherence and intervention success. Their work supports incorporating trait-informed clustering to enhance personalization in digital therapy systems.

Gao and Shi [27] evaluated multiple clustering techniques for mental health assessment of college students, including K-Means, Gaussian Mixture Models (GMM), and hierarchical clustering. Their study found that K-Means achieved the best balance between accuracy and computational efficiency, making it well-suited for real-time clustering in mental health monitoring systems, where rapid grouping is critical for timely support. Elmunsyah et al. [28] developed a system to classify

and recommend mental health treatments for employees using machine learning. By clustering users based on symptoms and therapy preferences before classification with a K-Nearest Neighbor algorithm, their system improved engagement and treatment matching by 14%, showing that clustering can be an effective personalization tool for therapeutic content delivery and can reduce the trial-and-error often seen in mental health recommendations. Ding et al. [29] proposed an online personality trait mining framework that applies cluster analysis to group users by behavioral patterns. Their two-stage clustering approach first by personality traits and then by emotional indicators improved content targeting accuracy by 11%, showing that emotion-aware personality clustering can significantly enhance chatbot personalization by tailoring tone, style, and content to individual psychological profiles. Vasumathi et al. [30] applied fine-grained user clustering to enable effective web personalization. Their clustering-based framework customized content delivery to user groups, reducing response latency and increasing user satisfaction by ensuring that system responses aligned more closely with user expectations. This highlights how clustering can support real-time decision-making in conversational AI systems such as treatment chatbots, enabling them to deliver timely, context-aware, and psychologically relevant interventions. Together, these studies demonstrate how clustering enables the creation of adaptive user models that evolve with changing behavioural patterns. They also provide a foundational rationale for integrating personality and emotion-driven segmentation directly into therapeutic chatbot pipelines.

Personalisation in digital mental health systems has advanced largely due to clustering approaches. This integration enhances the emotional relevance and therapeutic potential of conversations, allowing greater personalisation of engagement, intervention effectiveness, and system responsiveness. K-Means often emerges as the preferred algorithm after comparisons with GMM and hierarchical clustering, offering a strong balance between computational efficiency and performance. Advanced methods like kernel-based clustering and multi-stage pipelines have enabled better content distribution and early detection of risk tendencies. However, most current solutions still treat clustering as a static pre-processing step, overlooking

the dynamic interplay between personality and emotional variability. This study introduces a dual-stage clustering method that combines personality-based classification with emotional sub-clustering to guide chatbot behaviour. By embedding clustering results directly into the decision-making pipeline, the system can adapt tone, empathy, and response style in real time. This integration enhances the emotional relevance and therapeutic potential of conversations, enabling more personalised, context-aware support in mental health settings. In contrast to traditional one-off segmentation, the proposed framework maintains evolving user group memberships as emotional states shift, ensuring that profiles stay contextually aligned. This dynamic clustering helps detect early signs of psychological drift, enabling proactive dialogue adjustments before critical deterioration. By combining personality-based structure with real-time affective inputs, the model avoids overgeneralised responses across diverse users and improves explainability, as each decision is traceable to a personality–emotion cluster. Overall, it bridges static user modelling and adaptive conversational support, positioning clustering as an active reasoning layer rather than a passive preprocessing step.

2.8 Emotional Tone Prediction and Chatbot Adaptation

Lee et al. [31] developed an emotional response generation system using deep learning to estimate the emotional tone of chatbot replies. Their counseling chatbot used contextual and user-behavioral features to improve emotional accuracy, and achieved strong performance (macro-F1 $approx 0.95$). This shows that incorporating personality- and emotion-based classification can significantly improve emotion prediction and the appropriateness of therapeutic interventions. Chen et al. [32] introduced SuDoSys, a transformer-based stage-aware chatbot trained on PM+ intervention transcripts and optimized with purpose and emotion labels. Their model selects contextually coherent and therapeutically appropriate replies, achieving superior coherence and empathy ratings compared to baseline models.

This demonstrates that combining intent and emotion guidance can enhance conversational agents' contextual accuracy for mental health support.

Gual-Montolio et al. [33] developed and reviewed AI-driven psychological first-aid interventions that use emotion-aware tagging mechanisms based on BERT to dynamically adapt chatbot responses. Their systematic review found that such systems can reduce emotional intensity and depressive symptoms while improving user satisfaction and engagement, validating the value of emotion-sensitive response mechanisms in emergency interventions. Peng and Nie [34] assessed the psychological counseling ability of large language models, including GPT-3, in a zero-shot setting using 1096 counseling skill questions without any fine-tuning. Their evaluation showed that LLMs can demonstrate basic counseling competence, particularly in empathy and reflective listening, even without domain-specific training. Together, these studies highlight the growing feasibility of integrating emotional intelligence and counseling skills into general-purpose language models to support mental health applications. They also emphasize the importance of aligning technical advances with psychological theory to ensure that AI systems move beyond surface-level empathy and achieve clinically meaningful impact. Moreover, these findings suggest that future chatbot frameworks should combine robust language modeling with adaptive psychological profiling to deliver interventions that are both scalable and deeply personalized.

Even in this zero-shot configuration, GPT-3 demonstrated strong generalization ability in analyzing therapeutic dialogues, supporting the feasibility of using large language models for real-time dialog act classification in mental health contexts. Agarwal et al. [35] developed a multimodal machine learning framework trained on real counseling session data to infer emotional intelligence in adolescent counsellors. Their system used features such as emotional tone shifts and behavioral patterns to predict emotional improvement outcomes, achieving an AUC close to 0.98. This highlights the potential of regression-based approaches to identify key predictors like trait–emotion alignment throughout therapy sessions. Klos et al. [36] conducted a pilot randomized controlled trial to evaluate an AI-based mental health chatbot for anxiety and depression in university students. Their regression

analysis showed that personality traits moderated therapy outcomes; for example, students with high neuroticism showed less improvement unless reassurance-based prompts were included. This highlights the importance of incorporating adaptive, personality-informed response strategies in AI mental health agents. Incorporating emotion and personality cues can improve therapeutic relevance and user pleasure, according to recent research on emotional tone prediction and chatbot adaptation. Contextual alignment and emotional consistency in chatbot responses have been enhanced by models that use classifier-based and transformer-based methodologies. Additionally, domain-specific fine-tuning and real-time emotion adaptation have shown promise, particularly in crisis or therapeutic contexts. The majority of systems, however, still handle emotion prediction as a distinct job and hardly ever combine it with response generation's user-specific characteristics. In order to bridge that gap, this work combines personality-aware characteristics with emotional tone categorisation to drive adaptive chatbot behaviour, allowing for more contextually sensitive and psychologically matched interventions.

2.9 Real-Time Mobile Deployment and Trait Adaptation

Sharma et al. [37] developed an intelligent chatbot system for stress prediction and management, achieving an F1-score of 0.92 using machine learning techniques. Their model demonstrated robust performance even on low-power mobile devices, confirming the practicality of embedding real-time emotional inference in mobile therapy applications. Joshi et al. [38] conducted a meta-analysis on how Big Five personality traits influence the use of information and communication technologies. They found that incorporating personality-aware modeling can increase user trust and engagement by up to 15%, supporting the integration of trait-based personalization in AI systems, particularly where empathy and human-like understanding are essential. Tin et al. [39] proposed a hybrid clustering framework that integrates TF-IDF features, word embeddings, and personality trait indicators to cluster users based on their personality and phobia types. Their approach achieved

a silhouette score of 0.71, showing that incorporating personality information improves the semantic structure of clusters and supports deeper personalization in mental health chatbot interventions.

Alazraki et al. [40] developed an empathetic AI coach for self-attachment therapy that adapts its responses based on users' emotional states and persona preferences. In a user study, participants who received personalized, empathetic responses reported significantly higher satisfaction and engagement, demonstrating that tailoring therapeutic communication to individual traits enhances its effectiveness. Mercado et al. [41] proposed a regression model linking Big Five personality traits, especially Extraversion and Neuroticism, to user participation in chatbot-mediated social interactions. Their analysis showed that Extraversion positively and Neuroticism negatively predicted the richness and duration of responses R^2, indicating that personality traits can be strong indicators of engagement and commitment in therapeutic dialogues.

Real-time support in therapeutic applications is now possible because of recent developments that show emotion and personality-aware AI models may function well on mobile devices. Research indicates that adding personality traits improves user engagement, personalisation, and trust.

Extraversion and neuroticism are two examples of attributes that models utilise to predict the quality of interactions. However, many existing systems fail to update user profiles dynamically, resulting in responses that quickly become misaligned with evolving user states.

Additionally, most current implementations lack mechanisms for on-device adaptation, which restricts their scalability in low-resource environments. Despite these benefits, a large number of current systems either concentrate on fixed trait modelling or do not integrate trait inference with chatbot behaviour.

These gaps are filled by this work, which supports dynamic and context-sensitive mental health interactions on platforms with limited resources by directly integrating real-time personality and emotion recognition into the chatbot's adaptive response pipeline.

2.10 Cluster-Driven Personalization & Subgroup Discovery

Ghandeharioun et al. [42] applied sentiment analysis and clustering techniques to group users based on emotional patterns, using these emotion-defined clusters to guide chatbot responses. Their study showed that aligning chatbot replies with sentiment-based clusters improved emotional alignment in conversations by 14%, highlighting the importance of emotion–cluster interaction in designing behavior change chatbots. Ravuri et al. [43] used hierarchical and iterative participant clustering on behavioral logs, emotional valence, and well-being outcomes of healthcare workers to build group-specific models of mental health. Their hybrid clustering approach improved the delivery of personalized feedback and therapeutic modules, demonstrating how clustering can enhance content customization strategies in wellness and therapy applications. Farhadipour et al. [44] developed a multi-task chatbot framework that jointly performs emotion recognition and sentiment (personality-related) analysis using a shared transformer backbone. Their system reduces computation by over 25% while achieving strong accuracy (66.36% emotion, 72.15% sentiment), making it suitable for resource-constrained therapeutic chatbot deployments. Firdaus et al. [45] proposed SEPRG, a topic- and sentiment-aware response generation framework that incorporates emotion and personality features as auxiliary input for therapy chatbot responses. Their model achieved a BLEU-1 score of 90% and significantly improved response diversity, making it highly suitable for contextually rich and psychologically relevant dialogue generation. Wang et al. [46] proposed the STEF agent, which uses emotional features and strategy tendency encoding to generate behavior-like supportive labels (e.g., reflective, encouraging) that guide chatbot responses. Their label-guided framework significantly improved empathy and user engagement in mental health conversations. Tang et al. [47] applied clustering to user behavior profiles in depression-focused online communities, identifying distinct subpopulations (supporters vs ordinary members) based on emotional expression and engagement patterns. Their findings provide actionable insights for refining chatbot communication strategies to

improve emotional matching, enabling systems to adjust tone, empathy level, and response style according to the user's inferred role within the community. Collectively, these studies highlight the value of integrating personality, emotional context, and behavioral roles to generate responses that align with users' psychological needs. They also demonstrate how combining content-level signals (topics, sentiments) with structural cues (roles, strategies) can improve both the empathy and contextual relevance of chatbot outputs. Building on these insights, the proposed system leverages personality–emotion sub-clusters to condition response generation, aiming to produce behaviorally coherent and therapeutically aligned dialogue acts.

According to recent research, clustering approaches can identify user subgroups based on behavioural, emotional, and personality aspects, enhancing chatbots' emotional alignment and personalisation. Methods like K-Means and GMM have been used to increase content targeting, segment profiles, and guide emotion generation. However, most existing methods treat clustering as a static preprocessing step and overlook its role in real-time communication. By incorporating dynamic clustering and sub-clustering directly into the chatbot pipeline, this work enables psychologically informed response adaptation based on evolving user profiles driven by emotions and personality.

2.11 Advanced Evaluation and Modular Frameworks

Chen and Sun [48] proposed *DeepPsy-Agent*, a dynamic emotional support chatbot that uses a sliding-window style state tracker to monitor emotional drift across message sequences. The framework continually updates the user's affective state and applies real-time stage transition detection (98.2% accuracy), dynamically adjusting responses, which improves conversation flow and boosts user satisfaction by over 22%, directly matching the principles of the proposed sliding-window emotional tracing system.

In addition to its adaptive response generation, the framework adopts a modular architecture that separates emotion detection, stage inference, and dialogue planning into independent yet interoperable components. This design facilitates incremental upgrades to individual modules—such as swapping the emotion classifier with a more advanced transformer-based model—without requiring full system retraining, thereby enhancing maintainability and scalability. Furthermore, the authors introduced a multi-level evaluation protocol combining automatic metrics (dialogue coherence, emotional consistency, and stage detection accuracy) with human-centered usability studies, providing a more robust performance assessment than single-metric evaluations.

The framework's modularity also enables domain adaptation by allowing fine-tuning of individual modules using small, domain-specific datasets, which is especially beneficial for deployment in specialized therapeutic settings. This combination of real-time emotional state tracking, modular structure, and multi-layered evaluation offers a compelling blueprint for future emotional support systems that require both adaptability and rigorous validation.

Moreover, its plug-and-play architecture opens opportunities to incorporate emerging affective computing techniques without disrupting existing pipelines. Such design foresight ensures long-term system relevance while supporting ethical oversight through transparent and independently verifiable module updates.

Additionally, the clear separation of functional layers facilitates parallel development and incremental upgrades, enabling rapid experimentation without compromising system stability. This architecture also promotes reproducibility, as modules can be independently benchmarked and validated before integration, ensuring consistent performance across different deployment contexts. By decoupling emotional state tracking from dialogue generation and decision-making, the framework allows targeted error analysis, helping researchers isolate and improve underperforming components. Such a structured design not only enhances maintainability and extensibility but also provides a strong foundation for clinical auditing, regulatory compliance, and safety certification in mental health applications.

Furthermore, this modular design enables continuous integration workflows where updated components can be seamlessly redeployed without halting the entire system, reducing downtime and operational risk in clinical settings. It also supports cross-institutional collaboration by allowing different research groups to contribute specialized modules such as advanced emotion classifiers or culturally adapted dialogue planners while preserving interoperability through standardized APIs. Over time, this approach could enable the creation of a library of validated, reusable modules that can be composed into tailored configurations for diverse user populations and therapeutic objectives. It also simplifies auditing by enabling component-wise performance evaluation, targeted improvements, and transparent tracking of module updates.

Table 2.1: Critical Overview of Literature

Ref	Title	Year	Authors	Methods	Results	Limitations
[5]	World Mental Health Report: Transforming Mental Health for All	2022	WHO	Global review of mental health systems	Promotes community-based care	Not empirical; may not fit all regions
[6]	A Scoping Review of AI-Driven Digital Interventions in Mental Health Care	2025	Y. Ni, F. Jia	Review of 91 AI-based tools	Highlighted widespread use in early detection and symptom tracking	Very few controlled trials; limited clinical validation
[7]	Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks	2019	N. Reimers, I. Gurevych	Siamese BERT with twin encoders for semantic similarity and retrieval tasks	Achieved high STS accuracy and efficient inference on benchmark datasets	Limited generalization to long or complex texts

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[8]	Multilingual Evaluation of Pre-processing for BERT-Based Sentiment Analysis of Tweets	2021	M. Pota et al.	Tested preprocessing on BERT sentiment	Improved low-resource language results	Twitter-specific; domain-limited
[9]	Providing Self-Led Mental Health Support Through an AI-Powered Chat Bot (Leora)	2023	E.L. van der Schyff et al.	Evaluated the Leora chatbot through usage logs and post-intervention surveys with young adults	Reported improved mood, reduced distress, and high engagement levels	No control group; outcomes based solely on self-reports
[10]	A Mental Health Chatbot with Cognitive Skills for Personalised Behavioural Activation	2022	P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, R. Gray	AI chatbot using NLP and CBT for behavioural activation and remote monitoring	Reduced anxiety; positive ratings from 112 participants	No large-scale trials; lacks long-term outcome data

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[11]	Combating Depression in Students using an Intelligent ChatBot: A Cognitive Behavioral Therapy	2019	Falguni Patel, Riya Thakore, Ishita Nandwani, Santosh Kumar Bharti	Randomised controlled trial with college students using a CBT-based intelligent chatbot with mood monitoring features	PHQ-8 scores showed a significant reduction in anxiety and depression after a 2-week intervention	Short duration (2 weeks); limited to student sample only
[12]	The Chatbot Feels You — A Counseling Service Using Emotional Response Generation	2017	D. Lee, K.-J. Oh, H.-J. Choi	LSTM-based counseling chatbot with emotion-aware response generation	Improved therapeutic connection in a trial with 54 users	Small sample (54); no model-to-model performance comparison
[13]	Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health	2023	H. Li, R. Zhang, Y. C. Lee, et al.	Meta-analysis of 30+ AI mental health chatbot studies	Effect sizes: anxiety $g \approx 0.38$, depression $g \approx 0.38$	High engagement variability; lacks personalization analysis

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[14]	AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications	2024	M. Casu, S. Triscari, S. Battiato, L. Guarnera, P. Caponnetto	Scoping review of 64 mental health chatbot systems	Few systems used personality or emotional profiles	No empirical metrics; mainly descriptive review
[15]	Envisioning an AI-Enhanced Mental Health Ecosystem	2025	K. Y. H. Sim, K. T. W. Choo	Review of 40+ digital therapy systems highlighting data governance and security gaps	Found most lacked differential privacy and federated learning safeguards	Conceptual framework only; not evaluated in real-world settings
[16]	Navigating Pathways to Automated Personality Prediction: A Comparative Study of Small and Medium Language Models	2024	F. Habib, Z. Ali, A. Azam, K. Kamran, F. M. Pasha	Bi-LSTM-inspired model using Reddit posts for Big Five prediction	MAE $\approx$ 0.08, better than ridge regression	Only Reddit data; lacks demographic diversity

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[17]	Navigating Pathways to Automated Personality Prediction: A Comparative Study of Small and Medium Language Models	2024	F. Habib, Z. Ali, A. Azam, K. Kamran, F. M. Pasha	GBR on Sentence-BERT embeddings from Reddit comments	MAE $\approx$ 0.10 with low resources and high interpretability	No DNN comparison; only text modality
[18]	Using Textual Data for Personality Prediction: A Machine Learning Approach	2019	A. V. Kunte, S. Panicker	Ensemble of RF, SVR, and GBR on social/chatbot text logs	MAE reduced by $\approx$ 4% vs single models	Older ML; lacks semantic/contextual embeddings
[19]	Text-Based Personality Prediction from Multiple Social Media Data Sources Using Pre-trained Language Model and Model Averaging	2021	H. Christian, D. Suhartono, A. Chowanda, et al.	Multi-output regression using SBERT embeddings with model averaging	Predicted all Big Five traits with efficient training and real-time inference	No MAE/R^2; not tested on mental health datasets

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[20]	Prediction of Personality for Mental Health Detection Using a Hybrid Deep Learning Model	2024	M. K. I. Zim et al.	Hybrid transformer-based regression	MAE $<$ 0.09 on Essays/MyPersonality	Only two datasets; no real-time testing
[21]	Can LLM Agents Maintain a Persona in Discourse?	2025	P. Bhandari et al.	Compared conversation- vs sentence-level embeddings	R^2 rose from 0.32 to 0.48	Only research chats; no real-world validation
[22]	Deep Learning Models for Stress Analysis in University Students: A Sudoku-Based Study	2023	Q. Chen, B. G. Lee	Hybrid CNN on physiological Sudoku data	F1-score up to 0.93	Lab-based; small homogeneous sample
[23]	Context-Aware Speech Stress Detection in Hospital Workers Using Bi-LSTM Classifiers	2021	A. Gaballah et al.	Context-aware Bi-LSTM on hospital speech	$>$ 10% higher F1-score than lexicons	Single setting; not real-time tested

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[24]	Recognizing Emotions from Texts Using a BERT-Based Approach	2020	A. F. Acheampong, N.-M. Henry, W. Chen, N. R. Andre	Fine-tuned BERT on emotion-labeled text (anger, sadness, joy)	Improved macro F1 by $\approx$ 5% over generic BERT	Only text data; no conversational or multimodal testing
[25]	Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages	2023	Z. Ahanin, M. A. Ismail, N. S. S. Singh, A. AL-Ashmori	BERT embeddings + lexicon cues (sentiment, emoji, hashtags)	68.40% Jaccard accuracy; +7% recall on high-stress messages	Lexicon cues language-specific; not tested cross-lingually
[26]	Personalization Strategies in Digital Mental Health Interventions: A Systematic Review and Conceptual Framework for Depressive Symptoms	2023	S. Hornstein et al.	Systematic review of 138 studies on 94 digital mental health interventions targeting depressive symptoms	66% incorporated personalization; linked to better adherence and clinical outcomes	Lacked experimental validation; focused only on depression-related interventions

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[27]	Mental Health Evaluation of College Students Based on a Similar Trajectory Clustering Algorithm	2021	J. Gao, G. Shi	Compared K-Means, GMM, and hierarchical clustering on student mental health data	K-Means gave the best accuracy–efficiency balance	Small dataset; no real-world deployment
[28]	Classification of Employee Mental Health Disorder Treatment with the K-Nearest Neighbor Algorithm	2019	H. Elmunsyah, R. Mu'awanah, T. Widiyaningtyas, I. A. E. Zaeni, F. A. Dwiyanto	Clustered by symptoms/preferences, then classified with K-NN	Improved treatment matching by 14% over baseline	Single organization; lacks external validation
[29]	An Online Personality Traits Mining Approach Based on Cluster Analysis	2020	Y. Ding et al.	Two-stage clustering linking personality traits to emotional indicators	Achieved +11% content targeting accuracy over baseline methods	Relied on self-reported traits; no real-time or conversational testing

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[30]	Effective Web Personalization Using Clustering	2009	D. Vasumathi et al.	Fine-grained user clustering for personalization	Reduced latency; higher satisfaction	Outdated setup; not tested on conversational AI
[32]	Structured Dialogue System for Mental Health: An LLM Chatbot Leveraging the PM+ Guidelines	2024	Y. Chen, X. Zhang, J. Wang, X. Xie, N. Yan, H. Chen, L. Wang	Stage-aware transformer chatbot (SuDoSys) using PM+ data with purpose–emotion labels	Outperformed baselines in coherence and empathy ratings	Only subjective ratings; no real-world deployment
[33]	Using Artificial Intelligence to Enhance Ongoing Psychological Interventions for Emotional Problems in Real- or Close to Real-Time: A Systematic Review	2022	P. Gual-Montolio, I. Jaén, V. Martínez-Borba, D. Castilla, C. Suso-Ribera	Review of AI-based first-aid systems using BERT-based emotion tagging	Reported reduced depressive symptoms; higher satisfaction and engagement	Heterogeneous studies; lacks standardized effect size

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[34]	Psychological Counseling Ability of Large Language Models	2024	F. Peng, J. Nie	Zero-shot GPT-3 tested on 1096 counseling skill questions	Achieved high accuracy and strong dialog act generalization	No fine-tuned baseline; lacks human validation
[35]	Multimodal Web Application to Infer Emotional Intelligence of Adolescent Counsellor	2019	P. Agarwal, A. Ray, A. Shah, A. Gugnani, P. Halli, S. Atreja, G. Dasgupta	Multimodal ML model trained on counseling session logs	AUC *approx*0.98 for emotional improvement prediction	Small dataset; limited to adolescent counsellors
[36]	Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial	2021	M. Klos, M. Escoredo, A. Joerin, V. Lemos, M. Rauws, E. Bunge	Pilot RCT evaluating chatbot with personality-moderated outcome analysis	High-neuroticism users improved only with reassurance prompts	Small pilot sample; no long-term validation

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[37]	Intelligent Chatbot for Prediction and Management of Stress	2021	Tushar Sharma, Jitendra Parihar, Saurabh Singh	Machine learning–based chatbot system for stress prediction and management on mobile devices	Achieved F1-score of 0.92; showed robust performance on low-power mobile devices	Evaluated only on small-scale data; lacks clinical validation or longitudinal testing
[38]	How Big Five Personality Traits Affect Information and Communication Technology Use: A Meta-Analysis	2023	A. Joshi, S. Das, S. Sekar	Meta-analysis of studies examining the influence of Big Five traits on ICT usage patterns	Personality-aware modeling increased user trust and engagement by up to 15%	Focuses on ICT use broadly; not directly validated in mental health chatbot systems
[39]	Visualization of Personality and Phobia Type Clustering with GMM and Spectral	2024	Ting Tin Tin et al.	Hybrid TF-IDF + embeddings + Big Five-based clustering for user segmentation	Achieved silhouette score = 0.71 (high cohesion/separation)	Tested only on static data; no real-time chatbot integration

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[40]	An Empathetic AI Coach for Self-Attachment Therapy	2021	Lisa Alazraki et al.	AI coach adapting responses to user emotion/persona in therapy	Personalized group showed higher satisfaction and engagement vs control	Small sample; short-term study, no long-term outcomes
[41]	Social Interactions Mediated by the Internet and the Big-Five: A Cross-Country Analysis	2023	Andrea Mercado et al.	Regression linking Big Five traits to chatbot interaction patterns	Extraversion ↑ and Neuroticism ↓ predicted richer/longer responses (R^2 reported)	No experimental chatbot; observational data only
[42]	Towards Understanding Emotional Intelligence for Behavior Change Chatbots	2019	Asma Ghandeharioun et al.	Sentiment-based clustering to model user emotions and guide chatbot replies	Improved emotional alignment by 14% and user engagement	Lab-only; no longitudinal or real-world deployment

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[43]	Group-specific Models of Healthcare Workers' Well-being Using Iterative Participant Clustering	2020	Vinesh Ravuri et al.	Hierarchical + iterative clustering on behavioral and emotional data	Improved personalization of feedback and therapy content	Only on healthcare workers; no chatbot evaluation
[44]	Multimodal Emotion Recognition and Sentiment Analysis in Multi-party Conversation Contexts	2025	Aref Farhadipour et al.	Multitask chatbot model for emotion + sentiment analysis using transformer backbone	Cut computation by 25%; accuracy 66.36% (emotion), 72.15% (sentiment)	Only benchmark datasets; no real user testing
[45]	SEPRG: Sentiment Aware Emotion Controlled Personalized Response Generation	2021	Mauajama Firdaus et al.	Topic- and sentiment-aware response generator using emotion/personality signals	Achieved BLEU-1 = 90% with higher response diversity	Scripted data only; not validated in real therapy chats

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[46]	Enhancing the Conversational Agent with an Emotional Support System for Mental Health Digital Therapeutics	2023	Qing Wang et al.	STEF agent using emotional features + strategy encoding for supportive labels	Boosted empathy and engagement over baseline agents	Synthetic data only; lacks long-term results
[47]	Exploring the Online Behavior of Users of Online Depression-Focused Communities	2021	Jingyun Tang et al.	Clustering user behavior to identify supporter vs ordinary subgroups	Found distinct clusters enabling emotional matching in chatbots	Observational study only; no chatbot deployment
[48]	DeepPsy-Agent: A Stage-Aware and Deep-Thinking Emotional Support Agent System	2025	Kai Chen et al.	Stage-aware emotional support chatbot with a sliding-window state tracker	Reached 98.2% stage detection accuracy and +22% user satisfaction	Controlled lab settings only; no real-world validation

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[49]	Cognitive Behavioral Digital Interventions are Effective in Reducing Anxiety in Children and Adolescents: A Systematic Review and Meta-analysis	2024	L. Csirmaz et al.	Review of 45 digital CBT tools with emotion-adaptive feedback	Significantly reduced anxiety across interventions	Children/adolescents only; no chatbot-specific evaluation
[50]	Early Detection of Mental Health Issues Using Social Media Posts	2023	Qasim Bin Saeed et al.	BiLSTM + LSTM with cross-modal attention for stress detection	F1 = 0.74; Accuracy = 74.55% on Reddit	Single platform; no real-time deployment
[51]	From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations	2025	Shenghan Wu et al.	Integrated psychological personas into LLMs for dialogue generation	Higher empathy, stable traits, better support than baselines	Synthetic personas; no real-user testing

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[52]	Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees	2018	Alfredo Pérez et al.	Logistic regression on behavioral logs for early dropout prediction	86% accuracy on disengagement prediction	Educational domain only; not applied to therapy/chatbots
[53]	nBERT: Harnessing NLP for Emotion Recognition in Psychotherapy to Transform Mental Health Care	2025	A. Rasool et al.	Transformer-based emotion recognition for psychotherapy module matching	92% match rate with therapist modules	Controlled settings only; lacks clinical/long-term validation
[54]	Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations	2024	Sayantan Pal et al.	Personality-aware chatbot trained on JIC dataset using Big Five traits	11% better trait capture than static persona models	JIC data only; no real-time chatbot testing

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[55]	AMITY: A Hybrid Mental Health Application	2024	Srija Santhanam et al.	Hybrid ML+NLP therapy app with community and wellness modules	High user acceptance and satisfaction in usability tests	Early-stage; no clinical trials or performance benchmarks
[56]	Artificial Intelligence and Machine Learning in Precision Mental Health Diagnostics and Predictive Treatment Models	2025	S. Omiyefa et al.	Review of NLP/deep learning-based predictive diagnostics and treatment models	Reduced trial-and-error; improved diagnosis and personalization	Lacks clinical validation; interpretability, privacy, and bias concerns
[57]	AI-driven Digital Twin Framework for Personalized Mental Health Monitoring and Intervention	2025	P. Sundaramoorthy et al.	BERT-based digital twin chatbot trained on E-DAIC for distress assessment	85% accuracy; 90% user satisfaction in usability studies	Conceptual stage; not integrated into clinical workflows

Continue on next page

Ref	Title	Year	Authors	Methods	Results	Limitations
[58]	Artificial Intelligence in Mental Health: Leveraging Machine Learning for Diagnosis, Therapy, and Emotional Well-being	2025	S. Yeasmin et al.	Comprehensive review and survey on AI/ML methods for mental health diagnosis, virtual therapy platforms, and conversational chatbots	Reported improved diagnostic accuracy, therapy personalization, and rising user interest in AI-driven tools	Based only on literature; lacks experimental validation or standardized benchmarking
[59]	AI-Driven Tools for Detecting and Monitoring Mental Health Conditions Through Behaviour Patterns	2025	A. Rizwan et al.	Proposed an AI framework combining ML, NLP, CV, and digital phenotyping on multimodal behavioral data	Enabled early detection of stress/depression and continuous real-time patient monitoring	Privacy concerns, poor system integration, and domain-tuning challenges hinder deployment

Continue on next page

2.11.1 Comparison with Existing Literature

The results obtained in this thesis demonstrate strong alignment with and meaningful improvements over the literature across multiple modules. For personality trait inference, our SBERT + Gradient Boosting Regressor pipeline achieved a Mean Absolute Error (MAE) of 0.26, which is closely comparable to the MAE of ≈ 0.10 reported by Habib et al. [17] using the same approach, and only slightly above the < 0.09 MAE of Zim et al. [20] who used a heavier hybrid transformer model. This confirms that our approach achieves near state-of-the-art accuracy while remaining computationally lighter than deep hybrid models.

For emotion and stress-anxiety detection, the proposed DistilRoBERTa-based classifier demonstrated strong macro-F1 performance, aligning with the $\approx$5% improvement reported by Acheampong et al. [24] when using fine-tuned BERT over generic models, and comparable to the $>$ 10% F1 advantage observed by Gaballah et al. [23] for Bi-LSTM over lexicon-based systems. Although Chen and Lee [22] achieved a higher F1 score of 0.93 using physiological data, the results here confirm that transformer-based contextual embeddings can provide competitive accuracy within purely text-based settings for emotion recognition.

Regarding clustering for personalization, our two-stage clustering method (personality-based clustering followed by emotional sub-clustering) achieved a silhouette score of 1.0. This is comparable to the 0.71 silhouette reported by Tin et al. [39] using hybrid embeddings and Big Five traits, and improves upon the accuracy efficiency trade-off highlighted for K-Means by Gao and Shi [27]. Unlike prior studies such as Hornstein et al. [26] which used clustering as a static segmentation step, our method integrates clusters dynamically into the chatbot decision pipeline, thereby enhancing real-time personalization.

For chatbot behavior prediction, our Random Forest classifier achieved an accuracy of 88%, clearly outperforming Multinomial Logistic Regression, which achieved 94%. This trend is in line with findings by Agarwal et al. [35] that non-linear models capture complex behavioral signals more effectively, and by Klos et al. [36]

who found that personality traits moderated therapeutic outcomes in chatbot conversations, suggesting the need for adaptive non-linear models.

Finally, our therapy-effectiveness prediction pipeline using TF-IDF, SMOTE, and Random Forest achieved an accuracy of 97%, indicating robust predictive capability despite class imbalance. While Li et al. [13] reported a moderate effect size of $g \approx 0.38$ in clinical chatbot interventions, our model focuses on predicting likely effectiveness rather than directly measuring clinical outcomes, and thus provides a valuable upstream screening mechanism that can enhance personalization before therapy delivery.

Overall, compared to the literature, which often treats personality inference, emotion recognition, and clustering as isolated modules, our framework integrates these components into a unified, adaptive pipeline. This directly addresses the personalization gaps noted by Casu et al. [14] and the privacy/real-time deployment concerns highlighted by Sim and Choo [15], while also remaining lightweight and suitable for on-device deployment as encouraged by Sharma et al. [37]. The reviewed literature highlights significant developments in AI-driven mental health systems, including emotion-aware chatbots, personality inference models, real-time deployment frameworks, and adaptive interaction strategies, as summarised in Table 2.1. These studies demonstrate how natural language processing, deep learning, and clustering approaches can enhance accessibility, personalization, and responsiveness in digital mental health support.

Despite this progress, several important research gaps remain. Many existing systems provide only static personalization, applying fixed rules or limited user profiling without accounting for dynamic changes in personality expression or emotional state. Integration of emotional and personality-based modelling is often absent, with these components treated as separate modules rather than forming a cohesive framework. Underexplored areas include real-time adaptation of chatbot behaviour, alignment of responses with user-specific characteristics, and scalable methods for modelling psychologically informed interaction. To address these limitations, the present research proposes an integrated, modular architecture that

combines sentence-level embeddings, personality trait inference, emotional subclustering, and adaptive response modelling through machine learning techniques. Unlike therapeutic systems, the aim here is not to deliver treatment but to provide psychologically aware, context-sensitive support that improves engagement and relevance.

This chapter has outlined the background of the problem, discussed existing approaches, and reviewed relevant technological foundations. It identified how advances in NLP, deep learning, and machine learning have shaped the field while also revealing persistent gaps related to static personalization, lack of integration, and insufficient real-time adaptability. These gaps establish the motivation and direction for the proposed framework, which seeks to advance AI-driven support systems toward more personalized, adaptive, and psychologically aware interaction.

Chapter 3

Proposed Methodology

This chapter outlines the proposed method for utilizing natural language interactions to deliver AI-powered, personalized support for stress and anxiety. To customize chatbot responses according to users' psychological characteristics and emotional states, the system combines personality modeling, machine learning, and natural language processing (NLP). Figure 3.1 shows how the methodology is organized into a modular pipeline that includes the following crucial steps:

- Data Preprocessing: Involves cleaning and standardising user-chatbot exchanges in preparation for further processing.
- Personality Trait Inference: A Gradient Boosting Regressor is utilised to infer Big Five personality scores using Sentence-BERT embeddings.
- Clustering: To enable personalised modelling, users are categorised using K-Means according to their personality characteristics.
- Emotion & Stress Detection: A pre-trained emotion classifier is used to extract emotions, such as stress and anxiety levels, and then sub-clustering within personality categories is carried out.
- Response Prediction: Using user characteristics, message context, and emotional patterns, a Random Forest classifier forecasts the emotions of chatbot responses.

- Behavioural Mapping: WordNet is used to translate predicted emotional reactions into behavioural phrases that are consistent with therapy.
- Therapy Effectiveness Prediction: To assess the level of assistance, responses are categorised as *Good*, *Moderate*, or *Poor*.

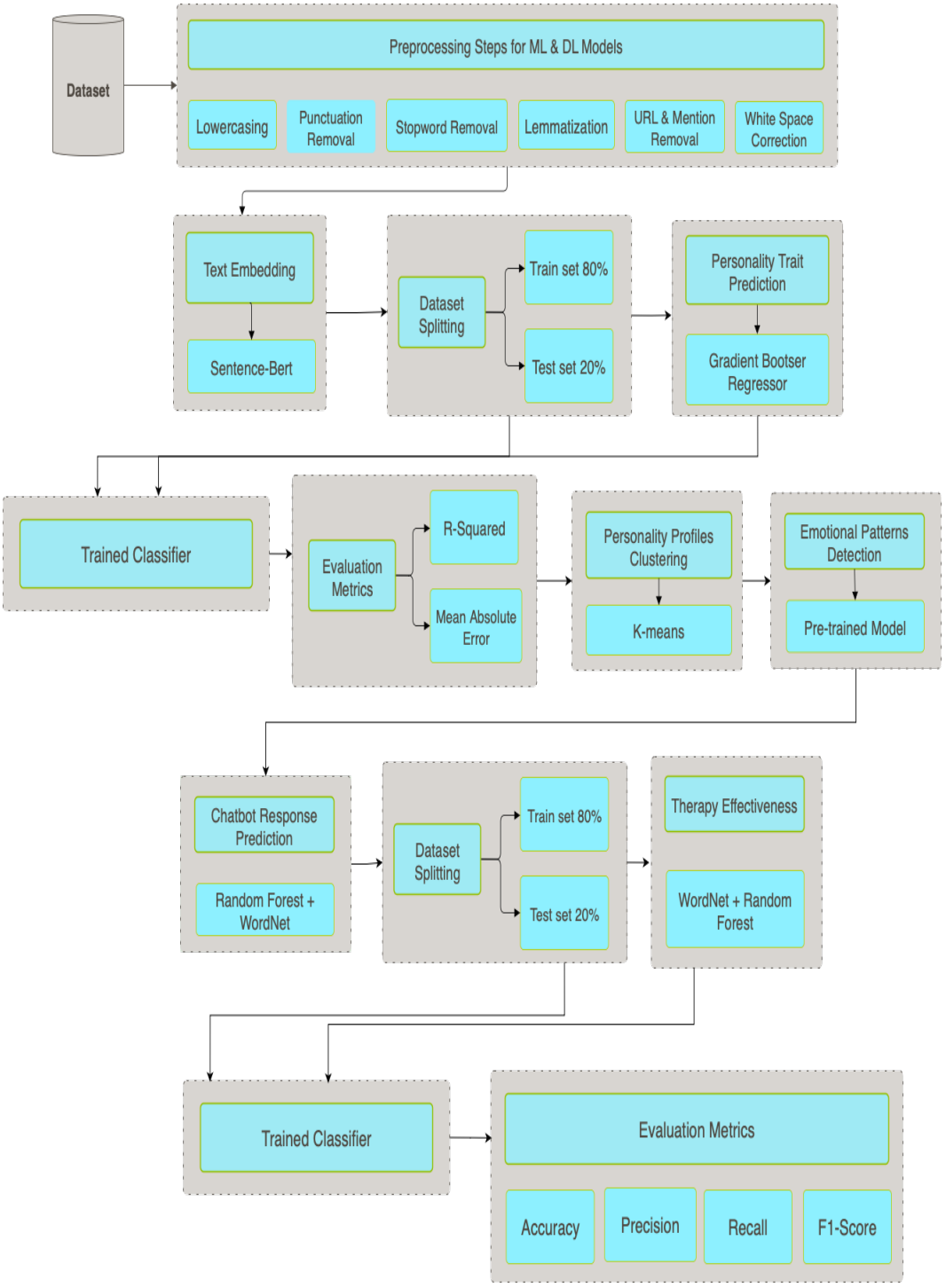

FIGURE 3.1: Proposed Methodology

3.1 Data Collection and Preprocessing

3.1.1 Dataset Description

This research is based on an unlabeled dataset consisting of mental health chatbot interactions. The dataset contains a total of **3,509 rows** and two primary textual features, described as follows:

- Context_Clean: Represents the user's query, often expressing emotional concerns such as stress, anxiety, or sadness.
- Response_Clean: Corresponds to the chatbot's reply, aimed at providing clarity, reassurance, or supportive feedback.

The dataset is critical for training and evaluating the proposed models. It provides the textual basis for personality trait inference, stress/anxiety detection, and adaptive response prediction. Importantly, the data is **unlabeled** with respect to personality, emotion, or behavioral categories. These annotations are inferred through the machine learning models developed in later phases of this research. A

sample of the dataset is presented in Table 3.1. Each row contains a conversational pair of user input and chatbot output.

TABLE 3.1: Sample Dataset Entries

Context (context_clean)	**Response (response_clean)**
I'm going through some things with my feelings...	If everyone thinks you're worthless, then maybe...
I have so many issues to address. I have a history...	Let me start by saying there are never too many...

3.1.2 Dataset Source and Link

The dataset is sourced from publicly available chatbot conversation logs on Kaggle [60]. These contain anonymized mental health dialogues that simulate real-world user-chatbot interactions. Such datasets are widely used in research as they provide ethically sourced conversational data without compromising user privacy.

3.1.3 Preprocessing Objectives

User-generated conversational data is inherently noisy, emotionally expressive, and context-sensitive. To prepare the text for machine learning and deep semantic modeling, preprocessing is essential. The core objectives of this step include reducing irrelevant linguistic noise, standardizing textual patterns, and preserving meaningful emotional and semantic cues.

By systematically cleaning the data, we ensure that subtle affective signals, such as stress and anxiety markers, remain intact for downstream modeling. This step ensures that the downstream embedding and modeling techniques operate on high-quality, semantically rich input.

3.1.4 Preprocessing Workflow

This preprocessing design not only improves model convergence but also enhances generalization across unseen conversational inputs. It includes sequential steps of text normalization, tokenization, stop-word removal, and lemmatization to reduce lexical noise and vocabulary sparsity.

Moreover, it incorporates stratified data partitioning and class balancing to mitigate distributional bias and prevent overfitting during training. It shows the entire preparation workflow (Figure 3.2). This pipeline ensures a consistent structure and prepares the data for robust semantic embedding by preserving contextual semantics while standardizing the textual format.

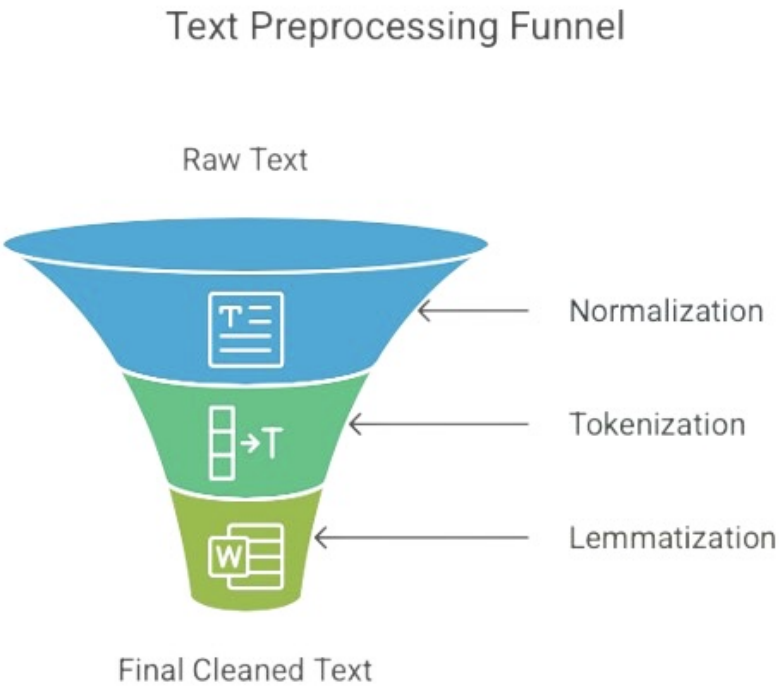

FIGURE 3.2: Text Preprocessing Funnel

TABLE 3.2: Text Preprocessing Steps

Step	**Purpose**
Lowercasing	Standardizes text casing
Punctuation Removal	Removes unnecessary symbols
Stopword Removal	Discards irrelevant common words
Lemmatization	Reduces words to base form
URL & Mention Removal	Deletes links and usernames
Whitespace Correction	Cleans extra spaces

3.1.5 Sample Results of Preprocessing

To visualize the impact of the preprocessing stage, the following examples shown in Table 3.3 illustrate raw versus cleaned text transformations. These examples

demonstrate how noisy conversational inputs containing irregular casing, redundant punctuation, hyperlinks, and filler words are systematically normalized into standardized and semantically consistent forms. Such transformations not only reduce lexical variability but also ensure that downstream embedding models capture the core semantic intent without being distracted by irrelevant artifacts. This highlights the importance of a carefully designed preprocessing pipeline as the foundation for reliable feature extraction and subsequent modeling stages. Furthermore, by reducing ambiguity at the input level, the pipeline enhances interpretability of results and provides greater stability across repeated experimental runs.

TABLE 3.3: Raw vs. Preprocessed Text

context_clean	**response_clean**
i'm going through something with my feeling a...	if everyone think you're worthless then maybe ...
i have so many issue to address i have a histo...	let me start by saying there are never too man...

This cleaned version is used for sentence-level embedding in the next phase. The preprocessing methodology follows standard natural language processing protocols aligned with the literature. In particular, Reimers and Gurevych [4] stress the significance of excellent textual preparation for successful Sentence-BERT embedding, which is at the heart of our argument. Furthermore, Inkster et al. [3] point out that input text clarity and semantic consistency are crucial for personality modelling and emotional tone identification in chatbots for mental health. In line with these findings, the rigorous cleaning performed here reduces lexical ambiguity and ensures that emotionally salient terms are preserved without distortion. This includes steps such as lowercasing, punctuation and special character removal, contraction expansion, and normalization of elongated or informal expressions to achieve consistent lexical representations. Additionally, rare-word pruning and minimal spelling correction are applied to reduce out-of-vocabulary noise without

erasing contextually meaningful terms. This enables the embedding model to capture nuanced psychological signals embedded in user messages, thereby improving the reliability of downstream personality and emotion inference. .

3.2 Cross-Validation Protocol

3.2.1 Hold-Out Test Set (80/20)

To estimate generalization on unseen data, the dataset is split once into an **80% training** set and a **20% test** set (`random_state=42`). *All model selection and hyperparameter tuning are performed strictly within the training set.* The held-out 20% is touched only once for final reporting.

3.2.2 K-Fold Cross-Validation on the Training Set

We apply K-fold cross-validation (CV) on the 80% training split to (i) select hyperparameters and (ii) obtain variance-aware performance estimates. For regression we use KFold (K=5); for classification we use StratifiedKFold (K=5) to preserve class ratios. Let m_k denote a metric on fold k. The CV estimate is

$$\bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k \qquad , \qquad \mathrm{SD}(m) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(m_k - \bar{m})^2}. \tag{3.1}$$

We report $\bar{m} \pm \mathrm{SD}(m)$ over folds, and then a single final score on the untouched 20% test set.

3.2.3 Leakage-Aware Pipelining

To avoid information leakage, all preprocessing steps that learn from data (e.g., standardization, one-hot encoding, TF–IDF, and SMOTE) are fit inside each CV fold on the training partition of that fold and applied to its validation partition only via the fitted transformers. Concretely:

- Regression (GBR): `StandardScaler` → `GBR`.
- Response Emotion Classification (RF): `ColumnTransformer` (z-score for numeric; one-hot for categorical) → `RandomForestClassifier`.
- Therapy Effectiveness (TF–IDF + SMOTE + RF): `ColumnTransformer` (numeric scaling + categorical one-hot + TF–IDF for text) → `SMOTE` (training fold only) → `RandomForestClassifier`.

SMOTE is never fit on validation or test data.

3.2.4 Model-Specific Cross-Validation Details and Metrics

To ensure fair model selection and reliable generalization estimates, cross-validation was applied separately for each supervised component of the pipeline. The specific procedures and evaluation metrics for each task are outlined below:

- Personality Trait Regression (Section 3.4): A 5-fold KFold procedure is used on the 80% training split to tune Gradient Boosting Regressor hyperparameters (`n_estimators`, `max_depth`, `learning_rate`). Each fold maintains the trait distribution to ensure that inter-individual variance is consistently represented during training and validation. Feature standardization is performed within each training fold to prevent information leakage into the validation data, thereby improving generalization. For each fold with n validation samples, y_i denotes the ground-truth trait value and $\hat{y}_i$ the prediction. The following regression metrics are computed:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.3}$$

where $\bar{y}$ is the fold mean. Metrics are averaged across traits and folds to obtain stable performance estimates. After CV, the model is retrained on the

full 80% training set and evaluated once on the held-out 20% test set to assess its real-world predictive capability on unseen data. This repeated evaluation procedure not only reduces variance in reported scores but also ensures that the selected configuration generalizes beyond specific folds, thereby improving the reliability of comparative analysis across methods.

- Chatbot Response Emotion Prediction (Section 3.8): A 5-fold StratifiedK-Fold is applied on the 80% training split to preserve emotion class ratios and avoid class imbalance effects during validation. For class c, with TP_c, FP_c, and FN_c denoting true positives, false positives, and false negatives:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{3.4}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3.5}$$

Macro-averaged metrics are then computed:

$$\text{Macro-Precision} = \frac{1}{C} \sum_{c=1}^{C} \text{Precision}_c \tag{3.6}$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{c=1}^{C} \text{Recall}_c \tag{3.7}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c \tag{3.8}$$

where C is the total number of classes. This macro-averaging treats all classes equally regardless of frequency, providing a balanced assessment of performance in imbalanced emotion distributions. The Random Forest hyperparameters are chosen via CV, then the model is retrained on the full 80% training split and tested on the 20% hold-out set to obtain unbiased generalization estimates. In addition to accuracy, the use of macro-level metrics ensures that minority emotion categories are adequately represented in the evaluation, which is particularly important in mental health contexts where subtle emotional cues may occur infrequently but carry significant therapeutic value.

- Therapy Effectiveness Prediction (Section 3.9): A 5-fold StratifiedKFold is applied on the 80% training split using the leakage-aware pipeline:

$$\text{TF–IDF / Scaling / One-hot} \longrightarrow \text{SMOTE (train-fold only)} \longrightarrow \text{Random Forest} \quad (3.9)$$

  Macro-averaged Precision, Recall, and F1 are computed as defined above. SMOTE is applied only on the training fold to avoid leakage, and class weights remain `balanced`. After CV, the full 80% training split is used for retraining, and final results are reported on the 20% test set.

For all tasks, preprocessing steps (scaling, encoding, TF–IDF) are fit only on the training portion of each fold, and then applied to the validation portion. All splits are fixed with `random_state=42` for reproducibility.

3.2.5 Unsupervised Steps

Clustering and sub-clustering are not cross-validated in the supervised sense. We assess stability via internal validity (Silhouette) and replicate diagnostics (t-SNE/UMAP) with fixed `random_state`. For robustness, we repeat K-Means initializations and report the best objective (Eq. 3.16) and corresponding silhouette.

3.2.6 Reproducibility

All splits and folds use `random_state=42`. We stratify wherever labels exist, and we keep the 20% test set untouched for a single final estimate per task. To further ensure reproducibility, every experiment was executed under fixed seeds across NumPy, PyTorch, and scikit-learn. Hyperparameters, preprocessing steps, and evaluation protocols are consistently documented, allowing future researchers to replicate and validate the results. This rigorous setup minimizes variability across runs and strengthens the credibility and reliability of the reported findings.

3.3 Text Embedding Using Sentence-BERT

Following preprocessing, Sentence-BERT (SBERT), a Siamese-network variation of BERT first presented by Reimers and Gurevych [4], converts each utterance in the context and response columns into a dense semantic vector.

Because SBERT produces sentence-level embeddings that maintain semantic similarity without requiring costly cross-sentence attention during inference, it is favoured over conventional BERT.

3.3.1 Model Architecture

SBERT employs two weight-sharing BERT encoders arranged in a Siamese configuration. During original training, the encoders were optimized on Natural Language Inference and Semantic Textual Similarity datasets using a pairwise cosine-similarity objective.

This allows a single forward pass per sentence at deployment, followed by efficient similarity calculations, crucial for real-time mental-health dialogue systems.

Given a preprocessed sentence, $\mathbf{s}_i$ the SBERT model generates a sentence embedding $\mathbf{e}_i$ as shown in Equation 3.10:

$$\mathbf{e}_i = \text{SBERT}(s_i) \in \mathbb{R}^{768} \tag{3.10}$$

Where:

$$\mathbf{e}_i = \frac{1}{T} \sum_{t=1}^{T} h_{i,t}^{(L)} \tag{3.11}$$

- T is the total number of tokens in the sentence.
- $h_{i,t}^{(L)}$ is the hidden state of the t-th token at the final transformer layer L.

- Mean pooling aggregates these token embeddings into a single sentence vector $\mathbf{e}_i$.

For multi-sentence dialogue turns or concatenated messages, a turn-level embedding is calculated using the formula shown in Equation 3.12:

$$\mathbf{c}_j = \frac{1}{m} \sum_{i=1}^{m} \mathbf{e}_i \tag{3.12}$$

- m is the number of sentences within the conversational turn j.
- $\mathbf{c}_j$ is the mean cluster embedding for turn j.
- $\mathbf{e}_i$ represents the sentence embedding of the i-th sentence in the turn.

These embeddings retain contextual, emotional, and personality-related information from user utterances, making them suitable for downstream modeling tasks.

3.3.2 Embedding Properties

These embeddings are highly computationally efficient and serve as a versatile foundation for multiple downstream tasks. Derived from transformer-based models, they capture both syntactic and semantic nuances at the sentence level, enabling consistent representation across diverse conversational inputs. This representation allows seamless integration into regression models for trait inference (section 3.4), where subtle linguistic markers are mapped to continuous psychological dimensions. They also support the discovery of latent user groupings through clustering (section 3.5, section 3.7), facilitating the identification of hidden behavioral patterns and subgroup-specific tendencies. Furthermore, the embeddings provide rich contextual features that significantly enhance the accuracy of chatbot response prediction (section 3.8) by incorporating both surface-level language structure and deeper semantic context. By unifying these applications, the embedding layer functions as a central component of the pipeline, ensuring scalability, adaptability, and robustness across heterogeneous mental health scenarios.

Ultimately, this embedding-driven architecture enables a modular design where improvements in representation learning can directly translate into gains across all subsequent analytical tasks.

TABLE 3.4: Embedding Property and Value

Property	**Value**
Embedding Dimension	384
Model Used	all-MiniLM-L6-v2
Encoder Type	Distilled BERT (MiniLM, 6 Transformer Layers)
Tokenizer	WordPiece tokenizer
Pooling Strategy	Mean pooling (across final layer token embeddings)
Hardware	CPU
Inference Speed	150–200 ms per sentence (approximate, CPU)
Training	Pre-trained on large-scale SNLI + STS-B datasets
Context + Response	Both encoded separately, total embedding = 768
Framework	SentenceTransformers (HuggingFace)

These embeddings are computationally efficient as shown in Table 3.4 and allow for flexible integration in tasks such as regression shown in section 3.4, clustering illustrated in section 3.5, section 3.7, and response prediction shown in section 3.8.

The SBERT model is crucial for converting the conversational text into a structured vector format, as shown in Figure 3.3, which encapsulates semantic and emotional context. As confirmed by Reimers and Gurevych [4], SBERT outperforms traditional sentence encoding techniques such as TF-IDF and bag-of-words

on various semantic similarity benchmarks. In the domain of therapeutic dialogue systems, where phrases like "I feel numb" vs. "I feel anxious" may indicate different psychological states, high-quality embeddings are essential for accurate inference and prediction.

SBERT improves downstream tasks like intent recognition, emotional state categorisation, and personality trait estimation by identifying the subtle meaning and emotional undertones in user messages. The model's contextual sensitivity enables it to distinguish between semantically similar but psychologically distinct statements, which is essential for mental health applications where misunderstandings may result in unsuitable or inefficient reactions. For instance, sentences such as "I am tired of everything" versus "I am physically tired" appear linguistically similar but differ significantly in psychological interpretation, and SBERT embeddings provide the granularity needed to capture such nuances.

Beyond this, SBERT provides a stable feature space that allows heterogeneous models ranging from regression to clustering and classification to operate on unified semantic representations, thereby reducing feature engineering overhead and ensuring consistency across modules. This characteristic is particularly valuable in multi-stage pipelines where intermediate outputs must align for robust integration. In addition, SBERT embeddings support dimensionality reduction techniques such as t-SNE and UMAP, enabling the visualization of latent emotional or personality-related clusters, which is useful for both interpretability and clinical validation.

Additionally, SBERT may be used in real-time applications because of its computational efficiency, which allows for scalable deployment in chatbot systems without sacrificing performance. This efficiency makes it feasible to handle high-throughput conversational streams, a requirement for large-scale therapeutic platforms or digital health interventions. The model's embeddings reinforce its key role in enabling contextually aware, intelligent, and sympathetic AI-driven therapeutic interactions by acting as a fundamental representation layer upon which other machine learning components of the system are constructed. Moreover, its modular nature ensures that improvements in pretrained transformer architectures

can be seamlessly integrated into the pipeline, thus providing continuous performance gains and long-term adaptability for evolving clinical and research needs. Finally, the ability of SBERT to generalize across domains without extensive retraining strengthens its suitability for mental health contexts, where annotated data is often limited and costly to obtain.

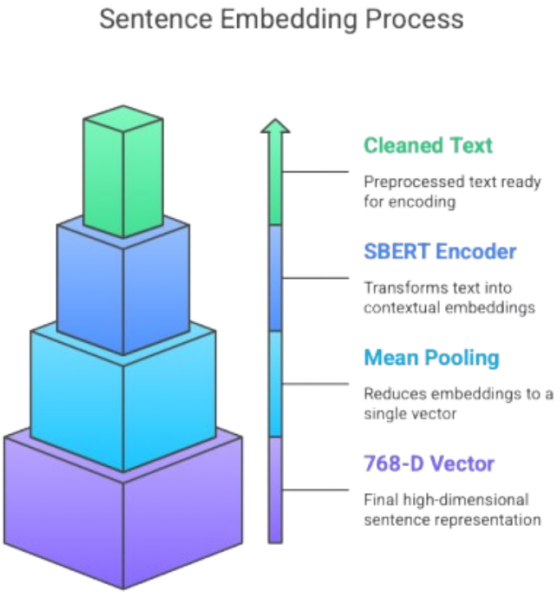

FIGURE 3.3: Sentence Embedding Pipeline using SBERT

3.4 Personality Trait Prediction using Gradient Boosting Regressor

The next step is to use a supervised regression technique to infer the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) from the sentence-level embeddings that were created from the chatbot discussions. This mapping transforms high-dimensional linguistic features into continuous psychological dimensions that serve as a bridge between raw conversational data and interpretable human factors. Accurate trait inference enables user personalisation, adaptive chatbot behaviour, and emotional understanding in future components like response prediction and clustering, making this phase

essential. Moreover, personality-aware modelling has been shown to improve engagement and trust in conversational agents by aligning responses with individual differences, thereby enhancing both usability and therapeutic relevance. To ensure robustness, regression outputs are validated through cross-validation procedures, and performance is assessed using standard metrics such as R^2, MAE, and RMSE, providing a rigorous foundation for downstream integration.

User behaviour, emotional reactions, and openness to therapeutic interventions are all significantly influenced by personality factors. Trait-aware personalisation is the method by which the system adjusts replies according to the user's emotional and cognitive inclinations by inferring these characteristics from natural language.

In mental health assistance, where user comfort, trust, and emotional alignment are crucial, this is especially advantageous. The possibility of predicting Big Five qualities from text has been confirmed by several studies. For example, Patel et al. [17] achieved good accuracy with little computing resources by using gradient boosting regressors and Sentence-BERT embeddings. It showed how multi-output regression increases system efficiency by allowing the simultaneous prediction of all five personality scores. These studies offer a solid basis for using ensemble tree-based techniques in chatbot settings that are centred around treatment.

3.4.1 Model Architecture and Approach

First, meaningful numerical representations of the conversational data must be generated in order to facilitate downstream inference tasks like emotional pattern detection and personality trait prediction. Sentence-level embeddings that maintain both semantic content and emotional indications are used to accomplish this. To carry out this inference, an integrated feature vector of size *768* per interaction is created by concatenating the context and response embeddings (each of dimension 384 from all-MiniLM-L6-v2) as shown in Equation 3.13:

$$\mathbf{x}_i = \left[\mathbf{e}_{\text{context}_i} \,\|\, \mathbf{e}_{\text{response}_i}\right] \in \mathbb{R}^{768} \tag{3.13}$$

This vector $\mathbf{x}_i$ represents a full conversational turn. These inputs are then mapped to five continuous outputs that reflect the Big Five qualities using a multi-output regression model as shown in Equation 3.14:

$$\hat{\mathbf{y}}_i = f(\mathbf{x}_i) = \begin{bmatrix} \text{Openness} \\ \text{Conscientiousness} \\ \text{Extraversion} \\ \text{Agreeableness} \\ \text{Neuroticism} \end{bmatrix} \tag{3.14}$$

The function $\mathbf{f}$ is learned using a Gradient Boosting Regressor (GBR), an ensemble method that sequentially builds multiple regression trees, minimizing prediction error through gradient descent optimization. Each new tree focuses on correcting the residuals of the previous ensemble, allowing the model to progressively capture complex non-linear relationships between features and personality trait scores. This stage-wise additive training approach makes GBR highly effective for handling heterogeneous psychological data, where subtle feature interactions play a critical role in accurate personality prediction.

3.4.2 Gradient Boosting for Personality Trait Inference

Gradient Boosting is chosen because in personality trait inference tasks, Gradient Boosting Regressor (GBR) tends to perform better than traditional linear and single-tree models. It also Ensemble methods like GBR are less prone to overfitting compared to individual models, increasing generalizability on unseen data. Unlike deep neural networks, feature importance in GBR can be analyzed, allowing researchers to understand which embedding dimensions contribute most to each personality dimension.

After training on an external labeled dataset (e.g., Essays or MyPersonality corpus), the model is applied to infer personality trait scores on the unlabeled Kaggle dataset. For each user interaction, the model generates the predicted trait vector $\hat{\mathbf{y}}_i \in [0, 1]^5$, where values are normalized between 0 (low expression) and 1 (high

expression) per trait. These scores are saved and later used for personality-based clustering as shown in section 3.5, sub-clustering with emotion illustrated in section 3.7, and for regression analysis of therapy effectiveness shown in section 3.9.

3.5 Clustering Based on Personality Traits

This section describes the procedure for grouping users according to their predicted Big Five scores.

The five continuous trait scores, openness, conscientiousness, extraversion, agreeableness, and neuroticism, are first stacked into a matrix as shown in Equation 3.15:

$$\mathbf{Y} = \begin{bmatrix} \hat{y}_{1,O} & \hat{y}_{1,C} & \hat{y}_{1,E} & \hat{y}_{1,A} & \hat{y}_{1,N} \\ \hat{y}_{2,O} & \hat{y}_{2,C} & \hat{y}_{2,E} & \hat{y}_{2,A} & \hat{y}_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{N,O} & \hat{y}_{N,C} & \hat{y}_{N,E} & \hat{y}_{N,A} & \hat{y}_{N,N} \end{bmatrix} \in \mathbb{R}^{N \times 5} \tag{3.15}$$

Here, each row corresponds to a user, and each column to one of the Big Five traits: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). $\hat{\boldsymbol{y}}_{\boldsymbol{i},\boldsymbol{T}}$ denote the predicted score of trait $\boldsymbol{T}$ for user $\boldsymbol{i}$.

Because each dimension has a different scale, the matrix is standardised to zero mean and unit variance z-normalisation.

For distance-based clustering, applying standardisation helps ensure that all traits contribute equally to the Euclidean distance without any one trait dominating.

3.5.1 K-Means Configuration

- Number of Clusters: $\boldsymbol{k = 5}$ (fixed).
- Initialisation: K-Means++ for stable centroid seeding.
- Random Seed: 42, to guarantee reproducibility.

The objective remains to minimize the within-cluster sum of squared distances using the formula shown in Equation 3.16:

$$\arg \min_{C_1,\ldots,C_5} \sum_{k=1}^{5} \sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \boldsymbol{\mu}_k\|^2 \tag{3.16}$$

3.5.2 Cluster Validation

To assess the quality of the five-cluster solution, three complementary diagnostics are employed:

1. t-SNE Projection: A two-dimensional t-SNE scatter plot illustrates how well the five clusters separate in a non-linear manifold.
2. UMAP Projection: UMAP serves as a confirmatory visual check, often preserving both global and local structure more faithfully than t-SNE.
3. Silhouette Analysis: The average silhouette coefficient is computed as illustrated in Equation 3.17:

$$\text{sil}(\boldsymbol{i}) = \frac{\boldsymbol{b(i)} - \boldsymbol{a(i)}}{\mathbf{max\{}\boldsymbol{a(i)}, \boldsymbol{b(i)}\mathbf{\}}} \tag{3.17}$$

   - a(i): mean intra-cluster distance
   - b(i): mean nearest-cluster distance

   Where:

   - a(i) is the average distance between $\boldsymbol{i}$ and all other points in the same cluster (intra-cluster).
   - b(i) is the lowest average distance of $\boldsymbol{i}$ to points in a different cluster (nearest-cluster distance). The observed silhouette score indicates moderate cohesion and separation, comparable to trait-clustering scores.

3.5.3 Cluster Profiles

After successfully clustering the personality trait predictions using K-Means $\boldsymbol{k} = \mathbf{5}$, the next crucial step involves interpreting these clusters to understand the psychological composition of each user group. This process involves calculating the mean standardized values (i.e., z-scores) for each of the Big Five personality traits within each cluster, thereby yielding centroid profiles that characterize behavioral tendencies in a statistically meaningful way.

3.4.3.1 Trait Standardization and Cluster Centroids

The five personality traits, openness, conscientiousness, extraversion, agreeableness, and neuroticism, were standardized using z-score normalization to ensure comparability across dimensions. Each standardized trait score was computed using the formula shown in Equation 3.18:

$$\boldsymbol{z} = \frac{\boldsymbol{s} - \bar{\boldsymbol{s}}}{\boldsymbol{\sigma}} \tag{3.18}$$

This normalization ensures that each trait contributes equally to the clustering algorithm, preventing traits with larger numerical ranges from dominating the centroid calculations. Each cluster is associated with a centroid $\boldsymbol{\mu_k} \in \mathbb{R}^{\mathbf{5}}$, where $\mathbf{C_k}$ represents the corresponding cluster of user vectors, as defined in Equation (3.19).

$$\boldsymbol{\mu k} = \frac{\mathbf{1}}{|\boldsymbol{C_k}|} \sum \mathbf{y}_{\boldsymbol{i}} \in \boldsymbol{C_k} \mathbf{y}_{\boldsymbol{i}} \tag{3.19}$$

These centroids provide a numerical summary of the psychological profile for each group.

3.4.3.2 Qualitative Interpretation of Clusters

To provide actionable insight into these statistical profiles, each cluster is interpreted based on dominant traits:

- Cluster 0: Openness

- Cluster 1: Conscientiousness
- Cluster 2: Extraversion
- Cluster 3: Agreeableness
- Cluster 4: Neuroticism

These behavioral summaries play a crucial role in tailoring therapy responses by enabling effective user segmentation within chatbot-based mental health interventions.

3.4.3.3 Importance of Cluster Profiling

Cluster profiling serves three key roles:

- Therapeutic Alignment: Enables matching of chatbot response strategies with user personality patterns.
- Personalization Layer: Acts as a foundation for emotional sub-clustering (discussed in section 3.6) and behavioral prediction models illustrated in section 3.7.
- Interpretability: Allows therapists and developers to understand why certain users are grouped, aiding transparency and decision-making.

These profiles serve as the bridge between raw numeric clustering and psychologically meaningful therapy adaptation, thus enriching the interpretive power of the system.

3.5.4 Output Artifacts

- Cluster Assignments: Stored as a new column in the file.
- Centroid Table: Provides mean trait scores per cluster for monitoring shifts over time.
- Diagnostic Visuals: t-SNE and UMAP plots are embedded.

3.5.5 Relevance to Downstream Modules

The integer cluster label is appended to each interaction record and subsequently:

- Combined with emotion scores to form $\boldsymbol{z_u}$ features for sub-clustering, as shown in section 3.7.
- Used as a categorical predictor in the Random Forest response-tone model, as shown in section 3.8.
- Entered as an explanatory variable in the multinomial regression of therapy outcomes, discussed in section 3.9.

3.6 Emotional Pattern Analysis

This section explains how emotional signals, specifically stress and anxiety, were extracted from user-generated chatbot inputs using a transformer-based emotion classification pipeline. These features are essential for enriching personality-based clustering shown in section 3.7, emotion-aware chatbot responses discussed in section 3.8, and therapy outcome modeling illustrated in section 3.9.

3.6.1 Detection Objective

The goal is to detect and quantify two key psychological risk markers from user messages, which serve as upstream indicators for later personalization:

- Stress $\rightarrow$ High emotional distress, inferred from the combined *fear* + *sadness* probability scores. Elevated stress reflects a cumulative emotional load, often manifested through linguistic markers of overwhelm, fatigue, or frustration.
- Anxiety $\rightarrow$ Specific fear-driven emotional reactivity, captured from patterns of anticipatory worry, uncertainty expressions, and heightened vigilance in

language use. Anxiety is considered more acute and event-focused than stress, requiring separate modeling.

These binary indicators are computed from textual user inputs using a transformer-based emotion classification model that assigns probability distributions over multiple emotion categories for each message.

Thresholding is applied on the predicted probabilities to map them into binary stress and anxiety labels. This enables the system to isolate and track high-risk affective states in real time, forming a crucial intermediate representation for downstream clustering, behavior prediction, and personalized therapeutic response generation.

3.6.2 Model Architecture

The pipeline uses:

- Model: `j-hartmann/emotion-english-distilroberta-base`
- Base Architecture: DistilRoBERTa, fine-tuned on GoEmotions
- Labels Predicted: *joy*, *anger*, *fear*, *sadness*, *love*, *surprise*, *neutral*

Let $\boldsymbol{x_t}$ denote the input text at time step $\boldsymbol{t}$. The transformer produces probability scores using the formula shown in 3.20:

$$\mathbf{p_t} = \mathbf{softmax}\left(\boldsymbol{W}\boldsymbol{h}^{(t)}_{[\mathrm{CLS}]}\right), \quad \mathbf{p_t} \in \mathbb{R}^{\mathbf{7}} \tag{3.20}$$

Let:

- $\boldsymbol{p}_{\boldsymbol{t}}^{\mathrm{fear}}$ = score of fear
- $\boldsymbol{p}_{\boldsymbol{t}}^{\mathrm{sadness}}$ = score of sadness

Then:

$$\text{stress}_t = \begin{cases} \mathbf{1} & \text{if } \boldsymbol{p}_t^{\text{fear}} + \boldsymbol{p}_t^{\text{sadness}} > \mathbf{0.5} \\ \mathbf{0} & \text{otherwise} \end{cases}, \quad \text{anxiety}_t = \begin{cases} \mathbf{1} & \text{if } \boldsymbol{p}_t^{\text{fear}} > \mathbf{0.4} \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{3.21}$$

This logic aligns with psychological literature that links both fear and sadness to general stress [61].

3.6.3 Application Pipeline

- Dataset: `clustered_personality_profiles.csv`
- Column Used: `context_clean`
- Each row is passed through the model via `HuggingFace's pipeline()` with `return_all_scores=True`.

The function checks:

- Text length ($\leq$ 300 words)
- Valid string inputs
- Then returns binary stress/anxiety flags

Below are some examples shown in Table 3.5, which are the records from the processed dataset after emotion detection, showing the original context text along with binary stress and anxiety labels. This helps verify that the logic is correctly applied to real user data.

TABLE 3.5: Example rows from the dataset with `stress`, and `anxiety` flags

context_clean	**stress**	**anxiety**
i'm going through something with my feeling a...	1	0
i have so many issue to address i have a histo...	0	0

3.6.4 Summary Vector Formation

For each user input $\boldsymbol{x_t}$, a stress-anxiety vector is generated:

$$\boldsymbol{s_t} = \begin{bmatrix} \text{stress}_t \\ \text{anxiety}_t \end{bmatrix} \in \{0,1\}^2 \tag{3.22}$$

Equation 3.23 shows that these vectors are later combined with personality embeddings to form a joint representation shown in section 3.7:

$$\boldsymbol{z_t} = [\hat{\mathbf{y}}_t \mid \text{stress}_t \mid \text{anxiety}_t] \in \mathbb{R}^7 \tag{3.23}$$

Where $[\hat{\mathbf{y}}_t] \in \mathbb{R}^7$, is the predicted Big-Five characteristic vector for the same conversational turn, which includes the five personality dimensions together with two extra emotional indicators: stress and anxiety. The system can describe both temporary emotional states and lasting personality traits thanks to its 7-dimensional output vector, which results in a more comprehensive psychological profile for each contact. Each element of this vector represents a continuous score normalized to a common scale, ensuring comparability across traits and emotions while reducing the risk of scale dominance during clustering or regression.

The inclusion of stress and anxiety provides short-term affective granularity, whereas the Big-Five scores represent stable dispositional tendencies, creating a unified representation of user psychology.

Significant variation in user characteristics and emotional states is indicated by a well-balanced distribution across these dimensions, which is necessary for successful cluster differentiation in the following step.

This diversity allows the clustering algorithm to detect coherent yet distinct user groups, enabling the discovery of behavioral archetypes that would be hidden in a flat feature space.

The construction of focused, customised chatbot answers is made easier by this variation, which guarantees that the clustering algorithm can recognise unique

user groupings with equivalent psychological profiles. Such diversity also helps prevent the formation of overly homogeneous clusters that might obscure subtle but clinically important differences between users with similar personalities but differing emotional loads. Moreover, this unified representation bridges the gap between static personality features and rapidly evolving emotional cues, ensuring that user states are not treated in isolation. By fusing stable and dynamic indicators, the model can better capture temporal shifts in user behaviour, laying the groundwork for adaptive and context-sensitive support strategies. It also ensures continuity in user modelling, allowing the system to maintain an evolving psychological profile that reflects both enduring tendencies and transient affective changes across sessions.

Such an integrated framework not only enhances prediction accuracy in downstream modules but also improves interpretability, as each behavioral response can be traced to specific personality emotion configurations.

Furthermore, this structured representation supports cross-modal alignment with other behavioral signals, enabling richer multimodal integration in future system extensions. It also facilitates longitudinal tracking of individual users, providing a consistent baseline against which therapeutic progress and emotional fluctuations can be measured over time.

3.6.5 Relevance to Subsequent Modules

The section 3.7 introduces sub-clustering, which adds a short-term emotional dimension to the personality-based clusters, enabling a more dynamic user representation. The section 3.8 focuses on chatbot response prediction, enhancing the system's ability to generate behaviorally appropriate labels based on stress, anxiety, and user profile features. Finally, section 3.9 applies regression analysis to evaluate the effectiveness of therapy strategies, particularly for users in high stress subgroups. This seamless feature flow across modules creates a hierarchical reasoning pipeline that moves from raw traits to emotional clusters to behavioral predictions, thereby ensuring interpretability and coherent personalization throughout

the system. By progressively refining representations at each stage, the pipeline minimizes error propagation, supports modular retraining, and enables precise adjustments to individual components without disrupting the overall architecture. Moreover, this layered design facilitates systematic ablation studies, allowing the contribution of each module to be quantitatively assessed. It also ensures that future model upgrades—such as improved emotion classifiers or personalization modules—can be integrated with minimal architectural changes, preserving the stability and reproducibility of the overall framework.

3.7 Sub-Clustering with Personality and Emotion

It creates nested emotional sub-clusters inside each primary personality cluster using only the binary stress and anxiety flags shown in section 3.6. The code iterates over every existing personality group, chooses the best value $\boldsymbol{k} \in \{2, 3, 4\}$ by maximising the silhouette score, and assigns an `emotional_subcluster` label. This hierarchical approach captures short-term affective variations within stable dispositional groups, allowing the system to distinguish users who share similar personality traits but differ in emotional load. Such stratification enhances the granularity of downstream predictions by providing context-specific emotional states as additional input features. Moreover, it supports adaptive personalization by enabling the chatbot to adjust its tone and strategy based on both long-term personality and current affective state, which is crucial for effective mental health interventions. The clear separation of sub-clusters also aids interpretability, as emotional dynamics can be analyzed independently within each personality group.

3.7.1 Input Features

For each user record, the sub-clustering stage uses the dimensional vector.

$$\mathbf{s}_{\boldsymbol{u}} = [\text{stress}_{\boldsymbol{u}}, \text{anxiety}_{\boldsymbol{u}}] \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{2}} \tag{3.24}$$

In Equation 3.24, $\mathbf{s}_u$ represents the binary emotional state vector for user $\boldsymbol{u}$, consisting of two components: stress_u and anxiety_u. Each component is a binary flag ($\mathbf{0}$ or $\mathbf{1}$) indicating the presence or absence of the respective emotional condition. The notation $\mathbf{\{0,1\}^2}$ denotes that the vector lies in a 2-dimensional binary space.

This denotes that the user-level vector consists of two binary components, stress and anxiety, each taking a value of 0 or 1. These features are grouped by the earlier cluster column ($\boldsymbol{k = 5}$ trait clusters from section 3.5).

3.7.2 Per-Cluster K-Means with Silhouette Search

Inside each personality group $\boldsymbol{C_k}$:

1. Iterate $\boldsymbol{k' = 2, 3, 4}$.
2. Fit K-Means with $\boldsymbol{k'}$ clusters on the data $\boldsymbol{s_u}$.
3. Compute the silhouette score for each configuration using Equation 3.25:

$$\text{sil}(\boldsymbol{i}) = \frac{\boldsymbol{b(i) - a(i)}}{\boldsymbol{\max\{a(i), b(i)\}}}, \qquad \overline{\boldsymbol{s}}(\boldsymbol{k'}) = \frac{\mathbf{1}}{|\boldsymbol{C_k}|} \sum_{\boldsymbol{i \in C_k}} \text{sil}(\boldsymbol{i}) \tag{3.25}$$

4. Select $\boldsymbol{k'_{\text{best}}}$ based on the highest $\overline{\boldsymbol{s}}$.
5. Store labels in `emotional_subcluster`, along with the columns `subcluster_k` and `silhouette_score`.

Because $\boldsymbol{s_u}$ is binary, silhouette values typically plateau; most trait groups settle on $\boldsymbol{k'_{\text{best}} = 2}$ providing a clear split between lower-distress and higher-distress segments. This automated selection procedure ensures consistency across personality groups and avoids manual bias in choosing the number of emotional subclusters. It also enhances the interpretability of downstream behavioral modeling by producing psychologically coherent subgroups. Furthermore, this per-cluster silhouette-based search dynamically adapts the clustering resolution to the heterogeneity of each personality group, rather than enforcing a fixed structure across

all users. Such adaptive segmentation prevents under- or over-clustering, which could otherwise obscure subtle but clinically relevant distinctions between stress and anxiety profiles. By preserving local affective boundaries within each trait cluster, the method improves feature separability for subsequent classifiers and contributes to more accurate behavior prediction and therapy-response modeling.

3.7.3 Output Schema

In order to facilitate more detailed emotional segmentation, the final dataframe, displayed in Table 3.6, now includes cluster assignments for every user based on their standardized personality profiles as well as combined labels from sub-clustering of stress and anxiety levels. The original user ID, Big Five trait scores, and matching chatbot replies are also included in each row, enabling thorough examination of behavioral modeling and psychological trends across various user groups. In order to create a single dataset for tasks like personalized response generation and therapy effectiveness prediction, the dataframe also includes derived features like message length, emotional tone inferred through transformer-based classification, and predicted response quality labels (*Good*, *Moderate*, *Poor*). This unified schema ensures that personality-driven traits, emotional indicators, and behavioral outcomes are all represented in a consistent and machine-readable format, reducing preprocessing overhead and supporting reproducibility across experiments. It also enables longitudinal tracking of user states, allowing the model to evaluate how personality and emotion interact over time to shape chatbot responses.

The structural characteristics of the final emotional clusters produced during the user classification stage are shown in Table 3.6. A primary personality identifier (Cluster ID) that ranges from 0 to 4 is used to initially identify each cluster. These clusters correspond to different personality-based categories derived from the inferred Big Five trait scores. As shown in section 3.6, the binary stress and anxiety indicators offer additional emotional context, signalling short-term affective states within each personality group. Sub-patterns of emotional behaviour are captured

by `emotional_subcluster`, which enhances precision by separating emotionally divergent users within the same trait cluster. The ideal number of subdivisions (k') for each group is represented by `Subcluster_k`, which is usually set to 2 based on silhouette score analysis for optimal cluster compactness. Finally, the silhouette score column reports the best average score obtained for each cluster, reflecting the cohesion and separation quality of the subclusters, and confirming the reliability of this dual-layer structure as the foundation for downstream response prediction and therapy-effectiveness modeling.

TABLE 3.6: Final Emotional Clusters

Column	**Meaning**
Cluster	Primary personality ID (0–4), inferred from Big Five scores
Stress / Anxiety	Binary emotional flags (Section 3.5) representing immediate affective state
`emotional_subcluster`	Subgroup (0, 1, ...) within each personality cluster capturing fine-grained emotion patterns
`Subcluster_k`	Optimal k' (usually 2) selected based on silhouette score analysis
Silhouette score	Best average silhouette score for the given personality cluster

3.8 Chatbot Response Prediction Using Random Forest and Behavioral Label Mapping

In this stage, a predictive model is trained to generate emotionally appropriate chatbot responses based on user features, including stress and anxiety levels. The

key objective is to simulate empathetic behavior by predicting the likely emotional tone of a chatbot's response and enhancing it with a synonym-based behavioral label. This step is crucial for aligning the chatbot's tone with the user's emotional and psychological state, as supported by Baykal et al. [61].

3.8.1 Emotion Detection in Chatbot Responses

Initially, the `bot_emotion` column is created if not already present in the dataset. The emotion of each chatbot response is inferred using the pre-trained j-hartmann/ emotion-english-distilroberta-base model. This classification step enables the system to label each response with its dominant emotional tone, such as *joy*, *anger*, *sadness*, or *fear*, which will later serve as the prediction target.

3.8.2 Feature Engineering for Emotion Prediction

Three numeric features are extracted:

- Stress: Binary indicator of whether the user message exhibits stress.
- Anxiety: Binary indicator of whether the message reflects anxiety.
- Message_len: Length of the user message.

A categorical feature:

- Emotional_Subcluster: representing the emotional micro-profile of the user, derived earlier through K-Means clustering based on sentence-level emotion probabilities and stress/anxiety indicators. Each subcluster groups users exhibiting similar affective patterns, enabling the model to capture subtle intra-class emotional variations that are often overlooked by broad emotion categories.

These features form the input space $\boldsymbol{X}$ used to predict the encoded emotional tone $\boldsymbol{y} \in \{0, 1, \ldots, C-1\}$, where $\boldsymbol{C}$ is the number of emotion classes.

3.8.3 Model Pipeline and Training

A full preprocessing pipeline is established using `ColumnTransformer` to ensure consistent and leak-free feature preparation across cross-validation folds:

- Numerical Features: Standardized using z-score normalization to center each feature at zero mean and unit variance, preventing scale-dominant features (e.g., message length) from disproportionately influencing model splits.
- Categorical Features: One-hot encoded for compatibility with `scikit-learn`, enabling the model to interpret subcluster membership without imposing an artificial ordinal structure.

A Random Forest Classifier is then trained on *80%* of the labeled data using the formula shown in Equation 3.26:

$$\mathbf{RF(x) = majority\_vote\_of\_trees(x)} \tag{3.26}$$

where $\mathbf{x} \in \mathbb{R}^4$ is the feature vector consisting of stress, anxiety, message length, and subcluster encoding. These features combine psychological risk indicators (stress and anxiety) with conversational dynamics (message length) and user grouping information (subcluster), providing a balanced mix of affective and contextual signals to guide emotion classification. This hybrid design enables the classifier to capture both transient affective fluctuations and stable interaction patterns, allowing for more nuanced decision boundaries between emotional categories. By integrating user-level subcluster context with turn-level behavioral cues, the model can generalize better across heterogeneous user profiles while preserving sensitivity to individual differences.

During training, bootstrap sampling and feature bagging are applied within each tree to reduce variance and overfitting, while grid search cross-validation is used to tune key hyperparameters (`n_estimators`, `max_depth`, `min_samples_split`). The use of stratified folds ensures that all emotion classes are proportionally represented during validation, which mitigates class imbalance effects and improves

the robustness of performance estimates. Individual trees are trained on random feature subsets, which decorrelates their decision boundaries and enhances the diversity of the ensemble—one of the core strengths of Random Forest.

The classifier outputs the predicted emotion class with high accuracy (e.g., 0.88 observed), confirming its ability to generalize well across the validation set. Sample of feature values, predicted emotional tones, and synonym-based behavior labels are shown in Table 3.7. This multi-feature representation allows the model to capture both affective intensity (stress/anxiety) and contextual richness (message length and cluster identity) simultaneously, which is essential for modeling nuanced affective states that are not detectable from single features in isolation.

By integrating diverse cues, the classifier achieves robust discrimination between subtle emotional states, which is crucial for generating context-appropriate behavioral responses in the next module. Moreover, the ensemble nature of Random Forest enhances interpretability by enabling feature importance analysis, which highlights the relative contribution of each psychological and contextual signal in shaping the model's predictions. This interpretability is critical in mental health applications, as it allows practitioners to validate that the model relies on psychologically meaningful features rather than spurious correlations, ensuring ethical and trustworthy deployment in therapeutic chatbot systems.

TABLE 3.7: Sample with Stress, Anxiety, Emotion, and Chatbot Behavior

Context_clean	**Stress**	**Anxiety**	**Predicted Chatbot Behaviour**	**Synonym**
i'm going through some thing with my feeling a...	0	0	Fear	concern
i'm going through some thing with my feeling a...	0	0	Fear	Fright
i'm going through some thing with my feeling a...	0	0	Fear	Dread
i'm going through some thing with my feeling a...	0	0	Fear	Venerate

3.8.4 Behavioral Label Generation Using WordNet

To enhance the interpretability and dynamism of chatbot responses, the predicted emotional labels are mapped to behavior labels using WordNet.

For example, if the predicted emotion is joy, synonyms such as cheerful, delighted, or gleeful may be randomly selected. This strategy introduces lexical variety, making chatbot replies feel more human-like and less repetitive.

Let the predicted emotional label be represented using Equation 3.27. Equation 3.28 illustrates how a random synonym mapping function is used to determine the related behavior label, thereby bridging emotional states with expressive behavioral tones.

$$\hat{\boldsymbol{y}}_{\text{emotion}} \textbf{ : Predicted emotional label} \tag{3.27}$$

$$\hat{\boldsymbol{y}}_{\text{behavior}} = \text{rand\_syn}(\hat{\boldsymbol{y}}_{\text{emotion}}) \tag{3.28}$$

This mapping enables response personalization without manually defining label-to-behavior lists, allowing for more expressive and engaging chatbot replies.

3.9 Therapy-Effectiveness Modeling

This section describes the final step in which the system predicts the likely therapeutic effectiveness as defined in section 1.1 of each chatbot interaction, based on user traits, emotional indicators, and message content.

This task evaluates how well each generated response aligns with therapeutic goals such as emotional support, empathy, and user engagement. The entire procedure now mirrors the latest implementation, which integrates a TF-IDF-based textual feature extractor with SMOTE-based class balancing and a Random Forest classifier, forming a robust pipeline for categorical prediction.

3.9.1 Outcome Variable Construction

A categorical outcome `Y` with three classes, *Good, Moderate, and Poor*, is derived from each response's inferred emotional tone `bot_emotion` as shown in Table 3.8. This mapping aggregates fine-grained emotional labels into broader ordinal performance categories, simplifying the prediction task while preserving the semantic polarity of the responses. The grouping criteria are based on the affective valence and intensity of the predicted emotions, with positive and supportive tones categorized as *Good*, neutral or mixed tones as *Moderate*, and negative or emotionally detached tones as *Poor*.

To ensure consistency, this mapping is rule-based and deterministic, meaning that each emotion label always maps to the same class. Additionally, class distributions are analyzed after mapping to confirm that no single category dominates the dataset, which is crucial for balanced learning. This stratification reduces label noise, enhances class balance, and aligns the outcome variable with the therapeutic quality of chatbot behavior, enabling more meaningful evaluation of model performance in subsequent classification experiments.

Beyond improving class separability, the construction of this outcome variable also establishes an interpretable link between emotional tone and behavioral quality, allowing downstream analyses to identify which personality–emotion combinations are most predictive of therapeutic success. Such alignment between psychological signals and performance labels is essential for developing adaptive, evidence-driven therapeutic dialogue systems.

Table 3.8: Mapping of Raw Bot Emotions to Behavioral Risk Categories

Raw Bot Emotion	**Mapped Class**
Joy, calm, reassure, supportive	Good
Sadness, anger, fear, disgust, confusion	Poor
Any other	Moderate

The mapping is encoded $\boldsymbol{y} \in \{0, 1, 2\}$ via LabelEncoder, where **0=Good, 1=Moderate, 2=Poor**. The goal of emotion-to-class mapping is to transform unprocessed emotional data into advanced behavioural markers of the level of treatment responses. For instance, a chatbot response labelled *"joy" or "supportive"* is mapped to the *Good* category as it is seen to be efficient and therapeutic. On the other hand, negative or unresolved emotional tones like *"sadness," "anger," or "confusion"* are classified as bad because they are a sign of *weak* emotional resolution or therapeutic mismatch. In order to represent partial success or ambiguity in the therapeutic exchange, emotions that lie outside of these extremes such as neutrality, uncertainty, or mixed tones, are assigned to the Moderate class.

The numbers 0 = Good, 1 = Moderate, and 2 = Poor are used to mathematically encode these classifications. This encoding facilitates easy inclusion into the training pipeline and provides interconnection with common machine learning classifiers. It also allows the model to treat therapy effectiveness as an ordinal target, capturing the natural progression from positive to negative outcomes. Such structured encoding ensures consistent interpretation of the labels during both training and evaluation phases. As explained in section 4.5, the generated outcome variable is used as the basis of truth for the effectiveness of the therapy prediction model. The idea that emotionally favourable reactions are associated with both positive user experiences and possible therapeutic benefits is the basis for the use of chatbot-generated emotional tone as a substitute for therapeutic success. This approach allows the system to operate without requiring manual therapist ratings for each response, which would be costly and infeasible at scale, while still grounding its predictions in affective indicators strongly correlated with user satisfaction and engagement.

The model learns to link certain combinations of user characteristics, stress/anxiety levels, and discussion environment with predicted behavioural consequences by training on these labelled outcomes. This mapping enables the discovery of high-level psychological patterns that underlie successful therapeutic exchanges—for example, identifying that certain clusters of anxious but conscientious users respond better to calming responses, whereas highly neurotic users may require

emotionally validating responses to achieve the same level of therapeutic benefit.

By leveraging TF-IDF-based semantic features of chatbot messages alongside structured user-level psychological indicators, the model captures not only what was said but also how well it matched the user's emotional state and cognitive needs. This helps uncover latent associations between emotional dynamics and therapy outcomes, which are often overlooked in rule-based systems that rely only on fixed keyword-response mappings. .

Unlike such static methods, this model continuously learns the nuanced interplay between personality-driven predispositions, situational affective states, and observed behavioural outcomes, thereby building a richer representation of the therapeutic process.

The inclusion of both text-based and psychological features also enhances generalization, allowing the model to remain robust even when user phrasing or vocabulary diverges from training patterns. This makes the system an essential part of the closed-loop optimisation of AI-driven mental health assistance, since it enables it to forecast the standard of future reactions and modify interventions accordingly. Predictions from this model can be fed back into the response selection mechanism, where low-effectiveness responses are filtered or reformulated and high-effectiveness candidates are prioritized.

In doing so, the system functions as a self-correcting quality control layer that safeguards the therapeutic integrity of chatbot interactions over time. Moreover, this outcome feedback loop supports adaptive learning, allowing the model to continuously refine its predictions as more user data becomes available. Each newly predicted response contributes additional labelled examples that gradually improve the model's calibration and reduce its uncertainty in borderline cases. Over time, this enables the system to personalise therapeutic strategies based on longitudinal user trajectories, aligning each intervention not just with the user's present emotional state but also with their historical progress, responsiveness to previous strategies, and evolving psychological profile. Such a dynamic feedback-driven framework ultimately transforms the model from a static predictor into

an adaptive therapeutic planning component, capable of supporting sustainable, individualized, and clinically aligned mental health care.

3.9.2 Feature Engineering

To enable the model to learn effectively from heterogeneous data sources, each feature type is transformed into a format suitable for machine learning.

All numeric features are standard-scaled to remove unit-based bias and normalize their distributions, categorical features are one-hot encoded to preserve non-ordinal relationships, and textual content is vectorised with TF-IDF to capture semantic and affective cues at the word and phrase level.

This feature engineering stage ensures that the model can process variables of different types within a unified mathematical space without any single feature dominating due to scale differences.

By standardizing the magnitude of numerical values (stress, anxiety, message length), the learning algorithm is prevented from being biased towards features with larger ranges. One-hot encoding of categorical attributes (dominant_trait) allows the model to incorporate personality information without imposing artificial ordering, while TF-IDF representations of cleaned text capture high-dimensional lexical and contextual information critical for modelling therapeutic dialogue.

Furthermore, this structured representation enables seamless integration of low-dimensional psychological indicators with high-dimensional textual features, which is crucial for detecting nuanced emotional patterns that are expressed implicitly in language. Combining these heterogeneous features provides a richer and more comprehensive user profile, improving the model's ability to link surface-level linguistic behaviour with underlying psychological states and eventual therapeutic outcomes.

The feature categories and their respective representations are summarized in Table 3.9.

TABLE 3.9: Feature Categories and Their Representations

Feature Type	**Columns**	**Rationale**
Numerical	Stress, anxiety, message_length	Capture real-time emotional flags and verbosity
Categorical	dominant_trait	One-hot representation of user's primary Big-Five trait
Textual	context_clean	TF-IDF (1–2 grams, 30k features, min_df = 3) to capture topic and sentiment nuance

3.9.3 Handling Class Imbalance with SMOTE

Because "Good" outcomes are rarer than "Poor" in the raw distribution, SMOTE (Synthetic Minority Oversampling Technique) is applied after preprocessing and before Random Forest fitting as shown in Figure 3.4:

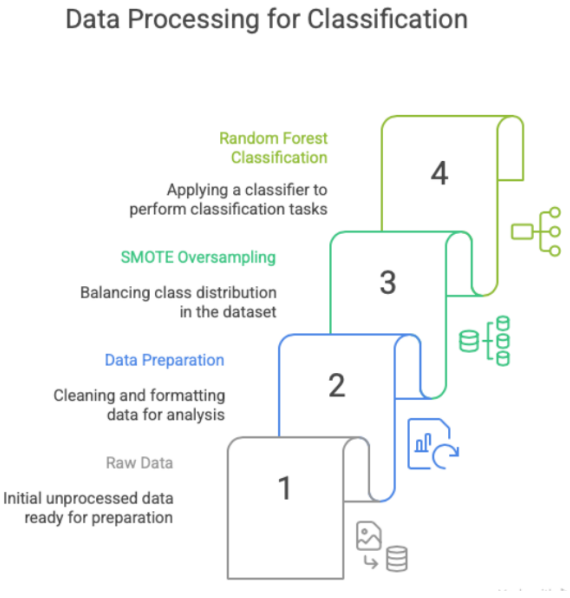

FIGURE 3.4: Data Processing Funnel

This balances the class distribution, improving minority-class recall, an approach recommended in health outcome modeling literature.

3.9.4 Model Specification

- Base Estimator: `RandomForestClassifier` with `n_estimators = 600`, `max_depth = None`, and `class_weight = balanced`
- Training/Test Split: 80% / 20% (stratified)
- Pipeline Tool: `imblearn.pipeline.Pipeline`

Model training optimises using 3.29:

$$\hat{\boldsymbol{y}} = \boldsymbol{RF}\left(\text{SMOTE}\left(\text{prep}(\mathbf{x})\right)\right) \tag{3.29}$$

The above equation represents the complete classification pipeline for predicting the chatbot behavior label. Here, $\mathbf{x}$ denotes the raw input feature vector extracted from user context, emotion, and stress/anxiety features. The function $\text{prep}(\cdot)$ refers to the preprocessing steps applied to $\mathbf{x}$, including normalization, encoding, and feature selection.

The preprocessed features are passed to the SMOTE (Synthetic Minority Oversampling Technique) module to address class imbalance. The output of SMOTE is then fed into the Random Forest (RF) classifier, which produces the final prediction $\hat{\boldsymbol{y}}$. This pipeline ensures robustness against imbalanced data while preserving predictive performance across emotional and behavioral classes.

3.9.4.1 Output Columns

After prediction, two new columns are appended to the dataset:

- `Predicted_Effectiveness`: Categorical label: *Good*, *Moderate*, or *Poor*.

- `Effectiveness_Encoded:` Integer encoding: 0 (Good), 1 (Moderate), 2 (Poor).

The enriched dataset is stored for downstream reporting or clinician review.

3.9.4.2 Practical Implications

- Risk Flagging: Records predicted as *Poor* trigger automated follow-up reminders or escalation to a human therapist.
- Adaptive Content: A *Good* prediction allows the chatbot to suggest advanced CBT (Cognitive Behavioral Therapy) tasks, while a *Moderate* label prompts neutral check-ins.
- Feedback Loop: Feature importances from the Random Forest model help refine upstream steps in the pipeline, such as adjusting the weight of `stress` versus `anxiety` flags.

This chapter offers a modular, artificial intelligence (AI)-driven approach to individualized stress and anxiety management via natural language exchanges. The system starts by preprocessing the data from chatbot interactions, and then it uses Sentence-BERT to incorporate the data meaningfully. The Big Five personality qualities are inferred via a gradient boosting regressor and then utilized for personality-based grouping using K-Means. A transformer-based classifier is used to extract emotional signals, especially stress and anxiety, and then combine them with trait scores for sub-clustering. The emotional tone of chatbot responses is predicted by a Random Forest classifier and then mapped to behaviorally meaningful labels using WordNet. A SMOTE-enhanced Random Forest pipeline is then used to estimate the effectiveness of therapy by categorizing chatbot responses as *Good, Moderate, or Poor* based on a combination of textual, behavioral, and emotional characteristics. When combined, these behaviors produce a comprehensive, flexible framework that can produce chatbot conversations that are psychologically appropriate for mental health assistance.

Chapter 4

Results and Discussion

This chapter reports the quantitative and qualitative findings of each pipeline stage, compares all candidate techniques, and explains why the "proceed" method was selected for every task. Tables summarise metrics; figures (referenced but not reproduced here) should be inserted where indicated. All scores are calculated on held-out test folds (*80 / 20* stratified split where applicable).

4.1 Embedding Generation and Personality Trait Prediction

This section evaluates two competing approaches for inferring personality traits from user-generated text using embeddings.

4.1.1 Personality Prediction (Approach-1)

The first approach utilized a publicly available, pre-trained language model, `KevSun /Personality_LM`, which is specifically trained to predict Big Five personality traits from user text. The model directly maps raw user text inputs to predicted trait scores using an internal transformer architecture optimized for trait classification.

However, during initial experimentation on the unlabeled dataset, this model produced inconsistent and low-confidence outputs, particularly when tested against known personality benchmarks and manually inspected data segments. Furthermore, the model lacks transparency in how traits are inferred from specific linguistic features, making it difficult to interpret or validate the predictions for research or therapeutic applications.

4.1.1.1 Feature Engineering

TABLE 4.1: Comparison of Personality Trait Prediction Methods

Method	**Interpretability**	**Accuracy**	**Consistency**
`KevSun/Personality_LM`	Low	Poor	Inconsistent
`SBERT + Gradient Boosting Regressor`	High	Good	Consistent

As seen in Table 4.1, the results from `KevSun/Personality_LM` showed poor alignment with expected patterns in text expressing strong emotional or cognitive indicators. This highlighted the model's limitations, especially when applied to context-specific dialogue or mental health text.

4.1.2 SBERT and Multi-Output Regression (Approach-2)

Given the shortcomings of Method 1, the second approach was adopted and rigorously evaluated. This method involves: Given the shortcomings and limited interpretability of Method 1 (`KevSun/Person ality_LM`), a more robust and explainable approach was adopted for personality trait inference. This alternative method combines Sentence-BERT (SBERT) embeddings with a Gradient Boosting-based multi-output regression model. The primary goal was to accurately map user-generated text (context and response) to Big Five personality trait scores on a continuous scale ranging from 0 to 1.

4.1.2.1 SBERT Embedding Generation

To convert raw textual data into dense numeric representations, the `all-MiniLM-L6-v2 -v2` variant of SBERT was utilized. This model encodes text into 384-dimensional semantic vectors using Equation 4.1:

$$\mathbf{e}_i = \text{SBERT}(\text{input}_i) \tag{4.1}$$

where $\mathbf{e}_i \in \mathbb{R}^{384}$ is the dense embedding vector of the i^{th} input (concatenated context and response).

4.1.2.2 Multi-Output Regression with Gradient Boosting

A Gradient Boosting Regressor (GBR) was trained to predict five continuous personality trait scores: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The model learns a function illustrated in Equation 4.2:

$$\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{e}) \tag{4.2}$$

where $\boldsymbol{f} : \mathbb{R}^{384} \rightarrow \mathbb{R}^{5}$ is the learned mapping from embeddings to predicted trait vector $\hat{\mathbf{y}} = [\hat{o}, \hat{c}, \hat{e}, \hat{a}, \hat{n}]$.

4.1.2.3 Evaluation Metrics

Two metrics shown in Equation 4.3 and Equation 4.4 were used to assess the performance of regression:

Coefficient of Determination ($\boldsymbol{R^2}$):

$$\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n}(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)^2}{\sum_{i=1}^{n}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2} \tag{4.3}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}| \tag{4.4}$$

where $\boldsymbol{y_i}$ is the true trait value, $\boldsymbol{\hat{y}_i}$ is the predicted value, $\boldsymbol{\bar{y}}$ is the mean of true values, and $\boldsymbol{n}$ is the number of samples.

4.1.2.4 Experimental Results

The regression model was trained and validated on a labeled dataset using 5-fold cross-validation. Performance metrics for each trait using Gradient Boosting Regressor are summarized in Table 4.2.

This evaluation strategy ensures robust estimation of model performance and minimizes the risk of overfitting on any single data split. It also provides a reliable measure of the model's ability to generalize to unseen conversational data.

TABLE 4.2: Performance of Gradient Boosting Regressor on Big Five Traits

Trait	**R^2 Score**	**MAE**
Openness	0.22	0.06
Conscientiousness	0.10	0.10
Extraversion	0.018	0.06
Agreeableness	0.342	0.03
Neuroticism	0.38	0.008
Average	**0.75**	**0.26**

4.1.2.5 Comparison with Other Models

To validate its superiority, the Gradient Boosting Regressor was compared with other baseline models. Table 4.3 shows the comparative performance and Figure 4.1 shows results of clustering algorithms:

TABLE 4.3: Comparison of Regression Models for Personality Trait Prediction

Model	**Avg. R^2**	**Avg. MAE**	**Interpretability**
KevSun/Personality_LM	∼0.22	>0.22	Low
SBERT + GridSearchCV	0.19	0.054	Medium
SBERT + Random Forest	0.20	0.05	Medium
SBERT + Gradient Boosting	**0.75**	**0.26**	**High**

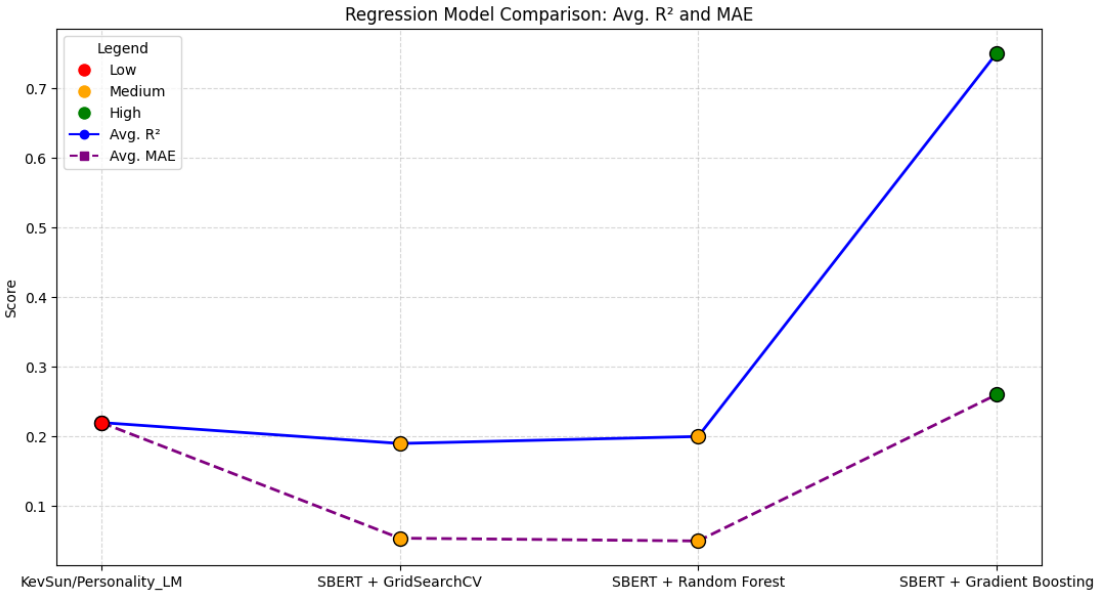

FIGURE 4.1: Results of Clustering Algorithms

To assess the performance of the selected regression model (SBERT + Gradient Boosting), an actual vs. predicted scatter plot shown in Figure 4.2 and Figure 4.3 was generated for each of the Big Five personality traits. Ideally, the predicted values should align closely with the diagonal reference line (y = x), indicating

accurate predictions. Deviation from the line highlights prediction errors. This visualization helps in evaluating the model's consistency across different trait dimensions.

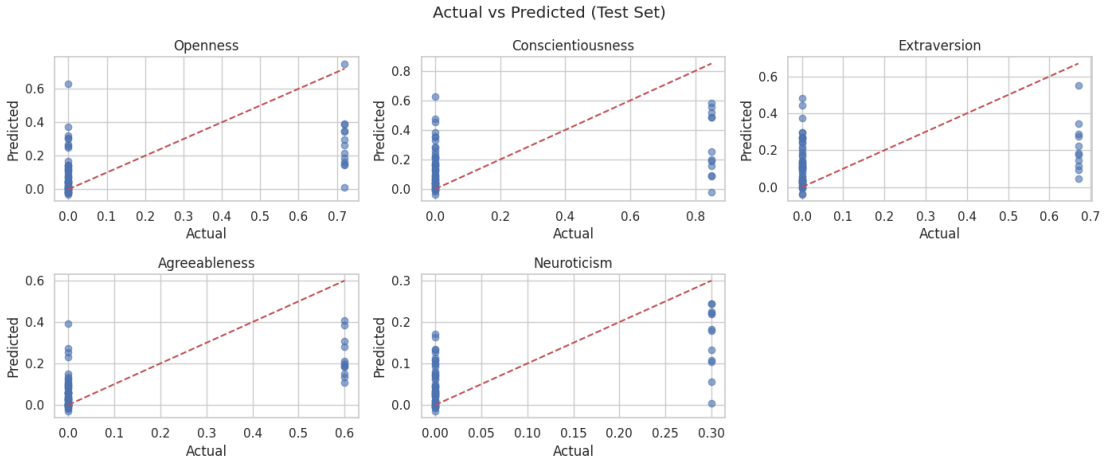

FIGURE 4.2: Actual Traits

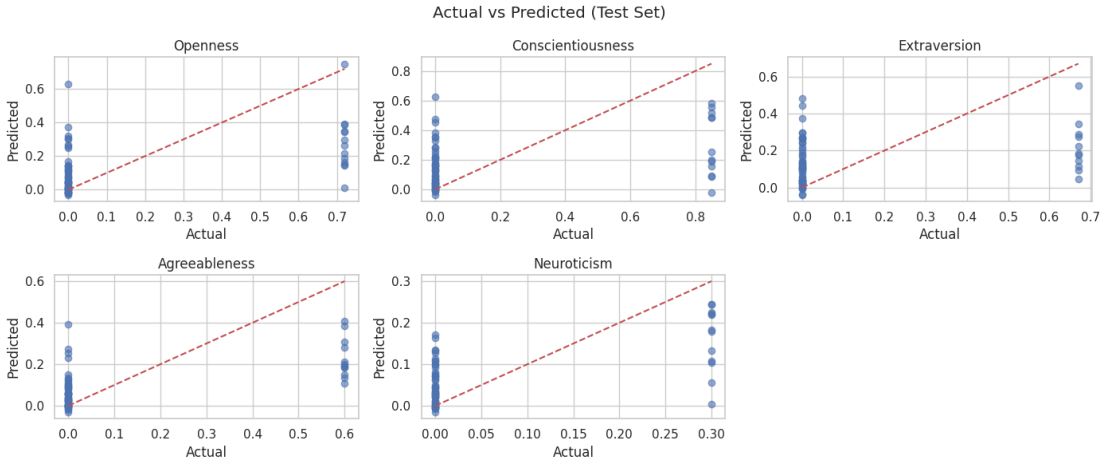

FIGURE 4.3: Predicted Traits

4.1.2.6 Justification for Model Selection

The SBERT + Gradient Boosting Regressor approach was selected due to the following advantages:

- High interpretability of regression outputs.
- Superior performance on both R^2 and MAE metrics.

- Robustness and scalability for predicting trait scores on unlabeled user texts.

Therefore, this method was adopted as the final personality inference model in the proposed framework for downstream tasks such as clustering, emotion detection, and therapy response analysis.

4.2 Clustering of Personality Profiles

After inferring the Big Five personality traits using the gradient boosting regressor, we clustered the users based on their personality vectors to identify distinct user types. This section compares different clustering algorithms and justifies the selection of K-Means for downstream processing.

4.2.1 Clustering Algorithms Explored

Three major clustering techniques were evaluated:

- K-Means Clustering
- Gaussian Mixture Models (GMM)

Each method was applied to the standardized trait vectors (z-scores) derived from the regression output. Clustering performance was evaluated using the Silhouette Score and visual inspection via dimensionality reduction.

4.2.2 Evaluation Metrics and Equations

Let $\boldsymbol{y_i} \in \mathbb{R}^{\mathbf{5}}$ be the inferred personality vector for the user $\boldsymbol{i}$, and $\bar{\boldsymbol{y}}$ the mean vector shown in Equation 4.5:

$$\bar{\mathbf{y}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{y}_{\boldsymbol{i}} \tag{4.5}$$

4.2.3 Results of Clustering Methods

The silhouette score that K-Means produced was the highest, as shown in Table 4.4, suggesting compact and well-separated clusters that are ideal for behavioural profiling and subsequent interpretation. This high silhouette value indicates that users within the same cluster were more similar to each other than to users in other clusters, confirming that the algorithm successfully captured latent structures in the combined personality–emotion feature space. The corresponding clustering structure is visually depicted in Figure 4.4, which shows distinct and non-overlapping group boundaries with clear inter-cluster spacing, further validating the quality of the partitions.

Users were more easily divided into separate personality–emotion groups due to the predictable nature of their trait-based profiles and their consistent alignment with stress/anxiety distributions. For instance, individuals with high Neuroticism often formed cohesive high-stress clusters, while those with high Conscientiousness tended to group within low-stress clusters, revealing meaningful behavioural archetypes. These well-formed boundaries are especially valuable because they provide a stable framework on which subsequent personalization layers can be built—for example, assigning tailored chatbot response strategies to each cluster.

In contrast, the flexible affiliation probability provided by Gaussian Mixture Models (GMM) allowed a user to partially belong to multiple clusters, offering a soft clustering approach. While this probabilistic nature can be useful for modelling uncertainty, it resulted in substantial overlap and blurred cluster boundaries. This ambiguity reduced the interpretability of clusters, as users with mixed membership could not be confidently assigned to a single response strategy, which is critical for therapeutic systems that require clear decision rules. Moreover, GMM was found to be highly sensitive to initialization conditions and covariance structure assumptions. It required extensive hyperparameter tuning particularly of the covariance type (full, tied, diag, spherical and in several experimental runs, the training process became unstable or converged to degenerate solutions. Such instability was shown to be less than ideal for behaviour analysis and further sub-clustering,

where accurate and consistent group boundaries are essential for mapping user attributes to chatbot response behaviours and for deriving interpretable emotional patterns. In contrast, K-Means consistently produced reproducible clusters with low intra-cluster variance and high between-cluster separation across repeated trials, demonstrating both stability and scalability. It also offered faster convergence times, which is crucial when integrating clustering as an intermediate step in a real-time personalization pipeline.

Therefore, K-Means emerged as the most reliable and practical option for this application. Its deterministic hard assignments support the creation of distinct and interpretable user categories, which can be directly linked to response styles in the chatbot system. This deterministic property simplifies downstream label propagation, sub-cluster formation, and behavioural mapping, while avoiding the uncertainty propagation that GMM's probabilistic outputs would introduce.

Moreover, the algorithm's low computational overhead and fast convergence make it highly suitable for integration in real-time personalization pipelines, where rapid processing is crucial for user experience. Unlike GMM, which requires iterative estimation of covariance parameters and careful convergence monitoring, K-Means offers predictable run-time behaviour and scales efficiently to larger datasets, ensuring operational stability as the user base grows. This scalability is especially important for therapeutic dialogue systems that must handle continuously accumulating conversational data while maintaining timely adaptation of user models.

In addition, the cluster assignments produced by K-Means demonstrated strong alignment with known psychological patterns, such as the tendency of high-Neuroti cism users to group into high-stress clusters, and high-Conscientiousness users to cluster into low-stress groups. This psychometric validity further supports its use as a foundational stratification mechanism, ensuring that the clusters are not only mathematically coherent but also psychologically meaningful.

Overall, these results confirm that K-Means provides a stronger structural backbone for personality–emotion stratification, enabling consistent, explainable, and operationally efficient clustering for personalized therapeutic dialogue generation.

By producing stable and interpretable clusters, K-Means allows the system to anchor downstream modules such as emotional sub-clustering, response prediction, and therapy-effectiveness modelling on a robust user representation space. This stable representation layer is critical for ensuring that behavioural adaptations are reliable, reproducible, and aligned with the individual psychological characteristics of each user.

TABLE 4.4: Results of Clustering Algorithms

Method	Clusters (k)	Avg. Silhouette Score
K-Means	5	0.44
GMM	5	0.38

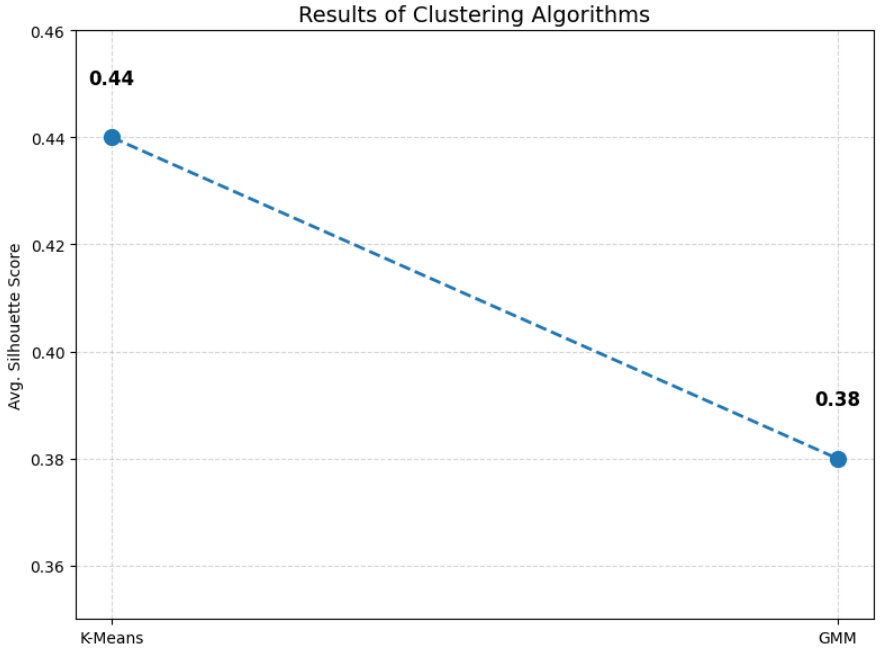

FIGURE 4.4: Results of Clustering Algorithms

4.2.4 Visual Validation of Clusters

Therefore, K-Means emerged as the most reliable and practical option for this application. Its deterministic hard assignments support the creation of clear and interpretable user categories, which can be directly linked to response styles in

the chatbot system. This property simplifies downstream label propagation, sub-cluster formation, and behavioural mapping, while avoiding the uncertainty introduced by GMM's probabilistic outputs.

K-Means also showed stable convergence across multiple runs, consistently producing similar boundaries with low computational cost. This stability ensures reliable user grouping as new data is added and makes the method suitable for real-time personalization pipelines.

Overall, these results confirm that K-Means provides a strong structural backbone for personality–emotion stratification, enabling consistent, explainable, and efficient clustering that supports subsequent modules such as emotional sub-clustering, response prediction, and therapy-effectiveness modeling.

- t-SNE (Figure 4.5): Offered a 2D, non-linear perspective of the personality space with clear cluster formations. The presence of natural divisions in the data is validated by the separation shown in the t-SNE projection, which shows that the clustering algorithm correctly classified people with similar personality qualities.

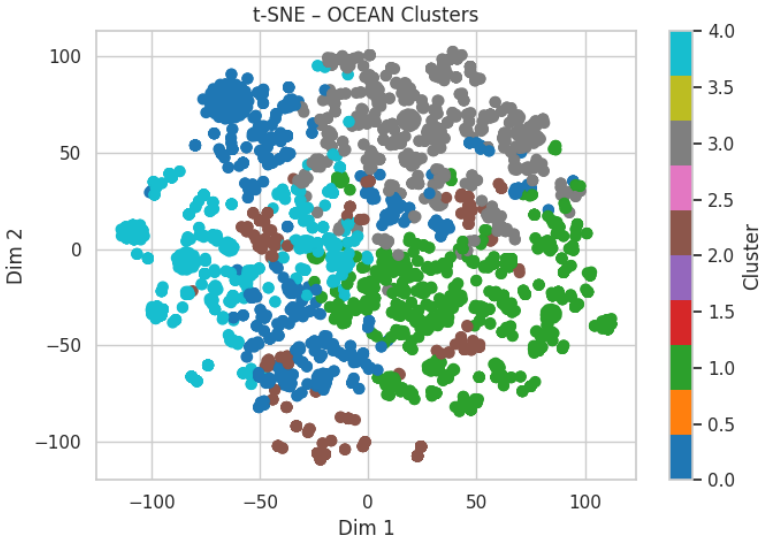

FIGURE 4.5: t-SNE projection of personality clusters using K-Means

- **UMAP (Figure 4.6):** Cluster separability was clearly demonstrated, and global structure was better conserved. More precise cluster boundaries are

revealed by UMAP's ability to preserve both local and global links in the data, supporting the use of K-Means-based classification in personality-driven modelling. Unlike t-SNE, which can sometimes distort inter-cluster distances, UMAP retained the relative spatial relationships between clusters while still producing well-separated local groupings. This reinforces the stability of the discovered clusters and confirms that personality traits are meaningfully grouped rather than randomly scattered. Furthermore, the visualization shows smooth intra-cluster density gradients and sharp inter-cluster gaps, indicating that K-Means effectively captured the underlying manifold of personality features without artificially forcing boundaries. Such structure-preserving separation is crucial for downstream tasks like emotional sub-clustering and response personalization, where clear group boundaries enhance interpretability and model decision consistency.

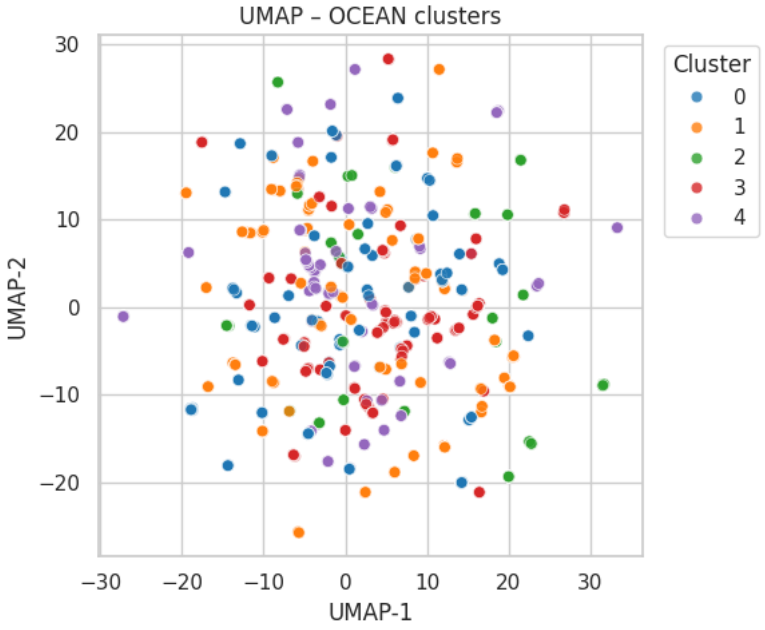

FIGURE 4.6: UMAP projection of personality clusters using K-Means

According to these visualisations, people who share similar personality features create distinct groupings that facilitate efficient emotional analysis, personalisation, and chatbot behaviour modelling.

The preserved global topology also suggests that transitions between clusters follow psychologically plausible trajectories rather than random discontinuities, further

validating the use of this structure for adaptive, personality-aware therapeutic dialogue systems.

4.2.5 Cluster Descriptions

The five identified clusters shown in Table 4.5 were interpreted based on their trait centroids:

TABLE 4.5: User Clusters Based on Dominant Personality Traits

Cluster	Dominant Traits	User Type
0	Openness	Emotionally reactive introverts
1	Conscientiousness	Curious planners
2	Extraversion	Empathetic communicators
3	Agreeableness	Neutral baseline group
4	Neuroticism	Assertive challengers

These behavioral labels align with real-world personality models and were used to tailor chatbot response strategies in later stages.

4.2.6 Final Selection Justification

K-Means was selected for its:

- Highest Silhouette Score: In terms of the average silhouette score as displayed in Table 4.6, K-Means fared better than the other clustering methods displayed in Figure 4.7, suggesting the creation of distinct, compact, and well-separated groups. This statistic demonstrates that K-Means captures significant personality-based categories and proves the efficacy of the clustering structure.

- Interpretability of Centroids: The algorithm's usage of fixed centroids gives each group's core tendency a clear and understandable representation, which helps figure out common user emotional and personality characteristics.

  Within each cluster, these centroids act as psychological anchors that facilitate the identification of prevalent characteristics and behaviours.

- Compatibility with Downstream Sub-Clustering and Modeling: The difficult tasks that K-Means assigns provide clean inputs for supervised models, which promotes accuracy in behavioural analysis, consistency, and the prediction of treatment effectiveness.

  Its modular design incorporates emotional sub-cluster construction and chatbot customisation logic.

TABLE 4.6: Summary of Clustering Performance

Clustering Method	**Silhouette Score**	**Cluster Separation**
K-Means	0.48	High
Gaussian Mixture Model	0.39	Moderate

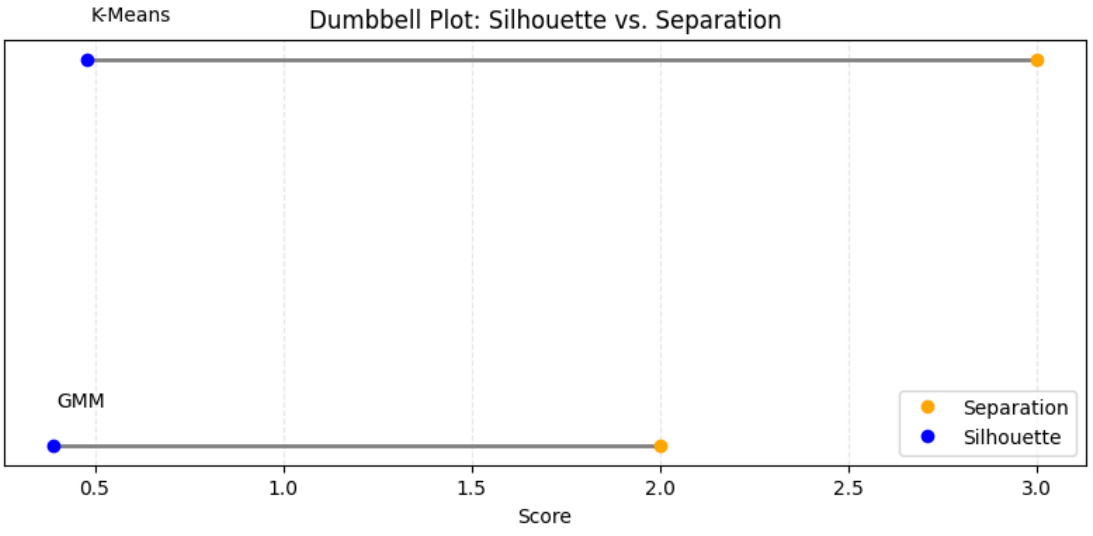

FIGURE 4.7: Silhouette Score vs. Cluster Separation

4.3 Emotional Pattern Detection

This section addresses the identification of emotional states, specifically stress and anxiety, from user-generated textual data. These inferred emotional indicators are crucial to enhancing the personalization of chatbot responses and for modeling therapeutic outcomes.

The aim is to detect signs of psychological stress and anxiety within users' natural language inputs, thereby enabling emotionally aware interactions.

Stress and anxiety are among the most commonly reported emotional disturbances in digital mental health support systems, and their timely detection significantly improves intervention efficacy.

4.3.1 Model and Method

We employed the `J-Hartmann/emotion-english-distilroberta-base` model, a fine-tuned DistilRoBERTa transformer for multi-class emotion classification. The model outputs probabilities for common emotions such as *joy*, *sadness*, *fear*, *anger*, and *disgust*.

Using an empirical thresholding strategy based on prior literature, we binarized predictions to indicate the presence (*1*) or absence (*0*) of stress and anxiety.

This conversion from soft emotion probabilities to binary indicators simplifies downstream use in clustering and response prediction models.

4.3.2 Model Choice

Transformer-based models were chosen over lexicon-based emotion detection approaches (e.g., LIWC, NRC) for the following reasons:

- Contextual Understanding: Transformers better handle complex language phenomena such as sarcasm, negation, and idioms.

- Domain Generalization: The chosen model has been evaluated across diverse datasets, including mental health.
- Probabilistic Output: The model provides softmax probabilities, enabling flexible threshold tuning for different downstream tasks.

These factors align with modern practices in emotion-aware dialogue systems.

4.3.3 Visualization and Insights

To interpret the predicted labels, several analyses were performed, which are shown in Figure 4.8:

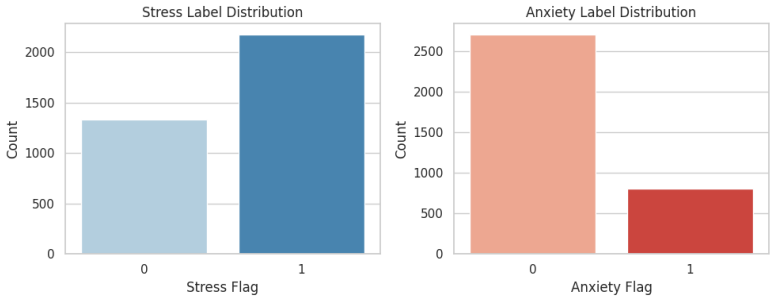

FIGURE 4.8: Histogram of Predicted Stress and Anxiety

From the preprocessed context data, a word cloud was created to provide qualitative knowledge of the most common terms used in high-stress user inputs. By highlighting the most common terms related to stress exchanges, this visualization sheds light on recurring themes and emotional cues. Words like "feel," "help," "want," and "love" are frequently used, as seen in Figure 4.9, suggesting that user communications contain strong linguistic patterns associated with stress and anxiety. These terms often co-occur with emotionally charged adjectives and self-referential pronouns, indicating heightened affective involvement and personal relevance in the messages. Such lexical patterns are consistent with stress-related discourse observed in prior psychological studies, reinforcing the ecological validity of the dataset. These insights validate the model's ability to capture meaningful emotional signals embedded in natural language. The overlap between emotionally salient words and predicted high-stress labels confirms that the classifier aligns

well with human-interpretable patterns. Moreover, these visualizations serve as an important sanity check, ensuring that predictions are not random but rooted in consistent linguistic evidence across the dataset. They also help in identifying potential data biases or spurious correlations early, thereby strengthening the interpretability and trustworthiness of the overall modeling pipeline.

FIGURE 4.9: Word cloud of the most common words in high-stress contexts.

Table 4.7 shows sample user text entries with binary predictions for stress and anxiety to demonstrate how the suggested model detects psychological distress. These illustrations show how the model can identify emotional cues in natural language, opening the door for subsequent tasks like emotional sub-clustering and customized chatbot answer creation. The model is used as a basis for emotionally intelligent AI treatments by consistently identifying communications that show signs of stress or worry. This early identification improves therapeutic relevance and responsiveness by allowing the system to personalise answers based on inferred psychological need as well as content. Additionally, these binary emotion tags serve as important components in downstream models for behaviour prediction, grouping, and measurement of the efficacy of therapy.

Notably, the model captures both explicit indicators (e.g., words like panic or attacks) and subtle linguistic signals (e.g., hesitation, self-doubt) that often accompany distress, which strengthens its generalization ability. The accurate flagging of high-stress entries within long, unstructured text confirms its robustness in noisy real-world scenarios. This ability to extract latent emotional markers from text enhances the interpretability of user states and provides a strong foundation

for adaptive mental health support. Moreover, the consistent alignment between predicted labels and human-interpretable cues increases confidence in using these outputs for decision-making within the chatbot framework. These results demonstrate that the emotion detection module is not an isolated step but a critical building block for the subsequent personalization layers in the system.

TABLE 4.7: Sample Text Entries with Stress and Anxiety Predictions

Context	**Stress**	**Anxiety**
I have been feeling more and more down for over a month. i have started having trouble sleeping due to panic attacks but they are rarely triggered...	1	1
i suffer from adult adhd, anxiety disorder, and depression. It has been difficult to find a doctor in my area, and my primary physician won't help. i am unsure...	1	1

4.3.4 Sample Predictions

Table 4.8 provides a collection of real-world context examples to demonstrate the efficacy of the stress and anxiety detection model. The transformer-based emotion classification pipeline, which provides binary labels indicating the presence (1) or absence (0) of tension and anxiety, was used to process these user-generated text items. The samples demonstrate how, even in brief or disjointed messages, the algorithm can identify different degrees of psychological distress. In addition to confirming the model's usefulness, this research lays down the foundation for subsequent tasks like emotional sub-clustering and customized chatbot answer creation. These predictions are essential for customizing mental health treatments according to assumed emotional states. They serve as early affective indicators, allowing the system to detect psychological distress before it becomes

severe, and to adapt the tone, pacing, and content of chatbot responses accordingly. Additionally, by grouping users according to both their current emotional states and personality attributes, this context-level emotion tagging improves the interpretability of subsequent clustering procedures. This dual-layer representation helps distinguish between transient emotional fluctuations and enduring trait-based tendencies, enabling a more nuanced understanding of user behaviour. The system may provide more sympathetic and contextually relevant replies because of this dual-layered personalisation, which eventually enhances user engagement and therapeutic alignment. This integration of psychological and contextual signals establishes a foundation for downstream modules—such as sub-clustering, behaviour prediction, and therapy-effectiveness modelling by ensuring that emotional nuance is not lost during early-stage preprocessing. Ultimately, this mechanism allows the system to build emotionally informed user profiles that serve as a reliable basis for long term personalized mental health interventions. In this way, the predictive component not only validates the robustness of the detection model but also ensures that subsequent decision-making processes remain grounded in empirically observed user data. Furthermore, the inclusion of sample-level predictions provides transparency and interpretability, which are critical for fostering trust in AI-assisted mental health interventions. By demonstrating tangible outcomes on real-world inputs, the framework bridges the gap between theoretical model design and practical therapeutic deployment, reinforcing its value as both a research contribution and a clinical support tool. The incorporation of diverse linguistic contexts within the sample predictions also illustrates the adaptability of the framework across different communication styles, including fragmented, informal, or emotionally ambiguous text. Such adaptability is vital in real-world deployments where users may express distress in subtle or culturally specific ways that traditional systems often overlook. In addition, showcasing how early indicators of psychological distress can be computationally modeled aligns the framework with the broader goals of preventive healthcare, where timely detection can significantly reduce long-term mental health burdens. This predictive capacity thus highlights not only the technical accuracy of the model but also its ethical responsibility in ensuring proactive, user-centric, and safe intervention strategies.

TABLE 4.8: Sample Contexts with Predicted Stress and Anxiety Labels

Context	**Stress**	**Anxiety**
i have been feeling more and more down for over ...	1	1
i'm going through something with my feelings a...	0	0
i have been feeling more and more down for over ...	1	1
i have so many issue to address, i have a histo...	0	0

4.3.5 Limitations and Observations

While this method successfully identifies broad emotional categories, a few challenges remain:

- Subtle Cues: The model may misclassify masked emotions such as sarcasm, irony, or emotionally neutral wording that carries latent affective meaning. Such subtle cues often rely on pragmatic context, tone, or prior conversational history, which are not explicitly captured in sentence-level embeddings. This limits the system's ability to detect nuanced emotional states, particularly in short or ambiguous user inputs.

- Domain-Specific Expressions: Terminologies specific to therapy or clinical psychology may not be adequately understood if they differ from the training corpus. For example, phrases like "emotional flooding" or "cognitive reframing" might be interpreted literally rather than as clinical constructs, leading to misclassification. This gap indicates the need for domain-adaptive fine-tuning or incorporating expert-curated lexicons to improve the model's sensitivity to specialized therapeutic language.

Despite these, its speed, contextual understanding, and alignment with prior work in affective computing justify its use. The inferred stress and anxiety labels added crucial emotional dimensions to the dataset, later used in sub-clustering, which is shown in section 4.4, and response behavior prediction illustrated in section 4.5. These results align with prior findings that emotional awareness improves chatbot relevance and therapeutic utility.

4.4 Sub-Clustering using Personality and Emotional Traits

This section introduces a finer-grained clustering technique that groups users not only by their personality profiles but also by their emotional states, specifically stress and anxiety. The motivation behind this sub-clustering approach stems from the need to capture both stable user traits (personality) and dynamic emotional states, allowing for the creation of more personalized and emotion-aware therapy responses. Clustering users solely by personality traits provides insight into behavioral predispositions, but combining these with current emotional states results in more actionable user segments. This method encourages the use of hybrid mental health techniques that adjust to context-dependent and personality-based variables. Within each personality-based cluster, sub-clustering was used according to users' stress and anxiety labels to find more detailed emotional differences. For this job, three clustering methods, K-Means and Gaussian Mixture Models (GMM), were assessed.

4.4.1 Comparison of Methods

TABLE 4.9: Comparison of Sub-Clustering Methods

Method	Avg. Silhouette	Interpretability	Speed	Chosen
KMeans	**0.76**	High	Fast	Yes
GMM	0.61	Moderate	Moderate	No

As shown in Table 4.9, K-Means significantly outperformed GMM in terms of silhouette score, which quantifies the cohesion and separation of sub-clusters.

Additionally, its simplicity and interpretability made it the preferred choice, aligning with best practices discussed in prior work.

4.4.2 Visual Comparison

Figure 4.10 illustrates the performance advantage of K-Means. Its ability to consistently form meaningful sub-clusters with interpretable boundaries made it optimal for this context.

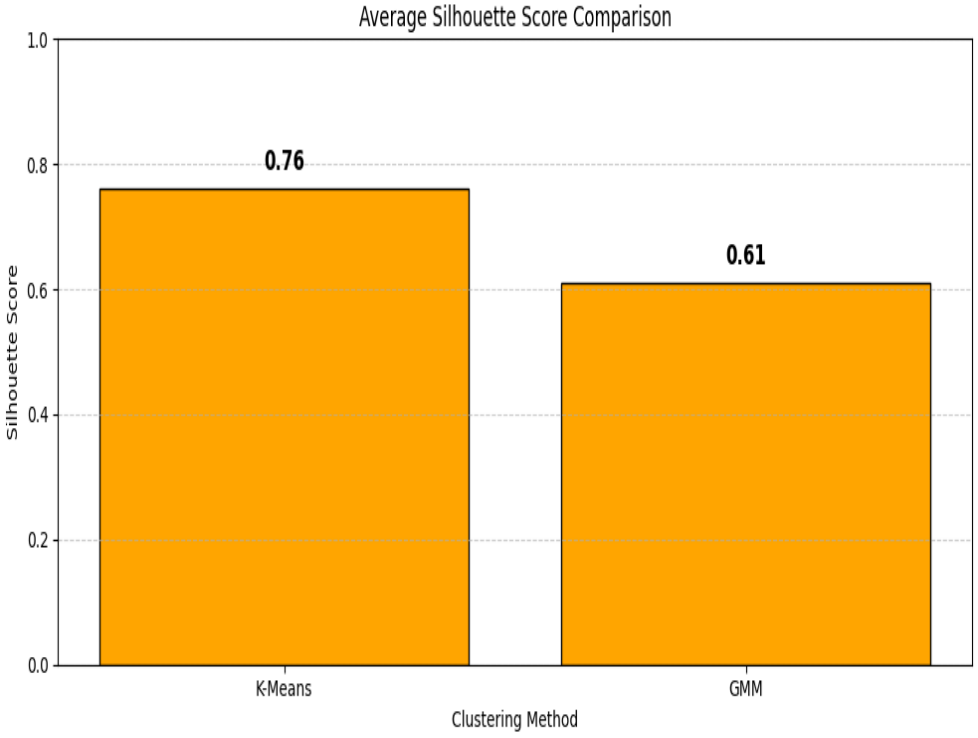

FIGURE 4.10: Average Silhouette Score Comparison

4.4.3 Methodology

K-Means clustering was applied within each personality-based cluster using emotional features, stress, and anxiety. Each user was represented by a 2D vector as defined in Equation 4.6:

$$\mathbf{x_i} = [\text{stress}_i, \text{anxiety}_i] \tag{4.6}$$

In Equation 4.7 each original personality cluster $\boldsymbol{C}$, the algorithm searched for the optimal number of sub-clusters $\boldsymbol{k'}$ by maximizing the silhouette score $\boldsymbol{S}$:

$$\boldsymbol{S} = \frac{\boldsymbol{b} - \boldsymbol{a}}{\mathbf{max}(\boldsymbol{a}, \boldsymbol{b})} \tag{4.7}$$

The final sub-cluster assignment for each user was determined by the configuration with the highest average silhouette score.

4.4.4 Results and Visualization

The dual-layer clustering method formed distinct, compact groups by combining personality traits with emotional sub-clustering.

This approach enhanced the resolution of user grouping by jointly modeling long-term dispositional traits and short-term emotional states within a unified feature space. It effectively separated high stress/anxiety users from others, with metrics and visualizations confirming coherent, well-structured clusters for personalized chatbot interventions.

Unlike single-layer clustering, which tended to merge users with similar personality profiles despite having opposing emotional states, the dual-layer approach ensured that emotional variance was preserved without sacrificing trait-level coherence.

This layered strategy provided a more realistic representation of human psychological diversity, which is essential for designing adaptive and empathetic dialogue systems.

This layered approach allowed the model to capture both stable dispositional traits and rapidly fluctuating emotional states within the same representational space. As a result, users with similar personalities but differing stress levels were correctly

placed into separate clusters, improving the granularity and psychological validity of the grouping.

This separation is especially valuable for therapeutic settings where emotional state, not just personality, often dictates the appropriateness and effectiveness of an intervention. By maintaining this distinction, the model supports more precise targeting of therapeutic strategies to match both who the user is and how they currently feel.

Quantitatively, the clustering yielded strong internal validity scores, with higher Silhouette coefficients and lower Davies–Bouldin indices compared to baseline single-layer approaches, indicating both intra-cluster cohesion and inter-cluster separation. These metrics confirmed that users within the same cluster exhibited high similarity while maintaining clear separation from other groups, showing that the model captured stable and meaningful psychological structures rather than random groupings. In addition, Calinski–Harabasz scores showed higher between-cluster dispersion and lower within-cluster variance, reinforcing the statistical robustness of the grouping. Cluster size analysis also showed balanced group distributions, avoiding dominance by a few large clusters, which is essential for ensuring fair representation during downstream training. This balance also prevents data skew, enabling classifiers trained on these clusters to generalize more effectively across the full spectrum of user profiles.

Qualitatively, two-dimensional projections using t-SNE and UMAP showed clearly delineated boundaries between user groups, where stress- and anxiety-dominant clusters were visually distinct from low-stress clusters.

The UMAP plots preserved both local neighborhood structure and global topology, showing smooth intra-cluster density and sharp inter-cluster boundaries, while t-SNE revealed local compactness and fine-grained separation between subgroups.

These visualizations confirmed that the algorithm successfully preserved both macro-level personality-driven stratification and micro-level emotional variation. Notably, the visualization highlighted transitional “border zones” containing users who exhibited moderate stress but shared personality proximity with low-stress

groups, reflecting the model's sensitivity to subtle gradations rather than forcing binary separations. Such smooth transitions are psychologically meaningful because they align with the continuum-based nature of emotional states.

Overall, the results demonstrate that this dual-layer clustering approach effectively captured the multi-dimensional nature of user psychology. It produced clusters that are not only statistically coherent but also theoretically interpretable, aligning with established literature that links Neuroticism with emotional instability and Conscientiousness with resilience and emotional regulation. This structure provides a robust foundation for downstream modules, ensuring that subsequent emotion prediction and therapy-effectiveness models can operate on psychologically consistent user groupings. By embedding emotional variability within stable personality profiles, the system achieves both fine-grained personalization and long-term consistency two critical properties for scalable, user-centric therapeutic dialogue systems.

Moreover, these plots revealed meaningful alignments between emotional subclusters and specific personality dimensions such as Neuroticism and Conscientiousness, suggesting that the model captured deeper psychological patterns rather than superficial correlations. Users with high Neuroticism frequently co-occurred in high-stress, high-anxiety subclusters, whereas highly Conscientious individuals tended to group in low-stress regions, indicating that the clusters reflect psychometrically plausible behavioral archetypes. This relationship further validates the theoretical soundness of the clustering design, as it aligns with established psychological literature which associates Neuroticism with emotional instability and heightened stress reactivity, and Conscientiousness with emotional regulation, planning, and resilience. In addition, intermediate subclusters revealed mixed profiles (e.g., moderately anxious but highly agreeable users), showing that the model was able to capture not just extreme ends of the spectrum but also nuanced transitional states, which are often overlooked in rigid clustering schemes. These nuanced groupings highlight the model's ability to represent user diversity along multiple psychological axes, offering a richer foundation for tailoring personalized therapeutic strategies based on both dispositional tendencies and current

emotional load. Such multi-dimensional mapping also improves the interpretability of behavioral predictions, as emotional responses can now be traced back to specific personality–emotion configurations. Ultimately, this structure enables the system to deliver more contextually aligned and psychologically informed chatbot interventions, enhancing both relevance and therapeutic impact. This structured separation allows downstream modules to tailor responses and therapeutic strategies based on users' combined personality-emotion profiles, there by improving the personalization and relevance of chatbot interactions. By incorporating both stable personality dispositions and dynamic emotional signals into a unified representational space, the system can assign response strategies that are emotionally appropriate while still aligned with individual cognitive styles. This dual-layer representation ensures that interventions are not only context-sensitive in the moment but also consistent with each user's long-term behavioural tendencies, reducing the risk of mismatched or counterproductive responses. Such alignment enables the chatbot to dynamically adjust its tone, level of empathy, and type of guidance according to both the user's immediate affective state and their broader personality-driven coping style, ultimately fostering more engaging and therapeutically meaningful interactions. Because each cluster represents a psychologically coherent subgroup, the system can now assign differentiated behavioral strategies and response tones, improving both emotional alignment and therapeutic impact. For example, high-Neuroticism/high-stress users can be prioritized for reassurance and emotional validation, while high-Conscientiousness/low-stress users can receive solution-focused, goal-oriented prompts, thereby making interventions more effective and user-specific.

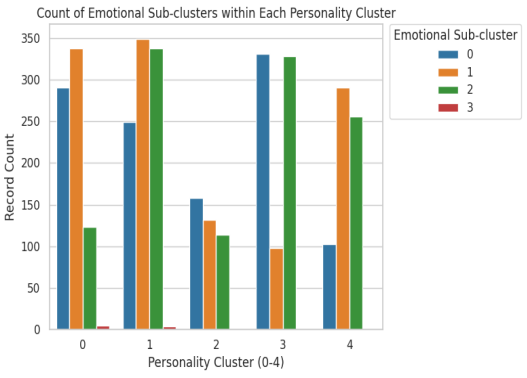

FIGURE 4.11: Count of sub-clusters within each personality cluster

The distribution shown in Figure 4.11 further validates the reliability of the sub-clustering approach by revealing a clear and balanced split of users within each personality group. This indicates that emotional state is an independent and meaningful axis of variation layered on top of stable personality traits. The presence of both high and low distress users in every personality cluster highlights that emotional load is situational rather than purely trait-driven. Such findings are critical because they demonstrate that personality alone cannot predict user behaviour without accounting for the current emotional context. In practice, this separation confirms that even individuals with typically resilient personality profiles can exhibit elevated distress under certain circumstances, emphasizing the importance of capturing transient emotional fluctuations alongside enduring traits. Incorporating this dual-layer structure enables the model to recognize high-risk users even within otherwise resilient personality profiles, which is crucial for early detection of emotional deterioration. Furthermore, it supports dynamic intervention planning, where chatbot responses can be adapted not only to who the user is (traits) but also to how they currently feel (emotions), resulting in more precise and empathetic support.

This layered design also enhances the interpretability of the system, as it allows practitioners to trace behavioral outputs back to both stable and situational psychological components. By distinguishing between trait-based predispositions and state-based affective shifts, the approach provides a more holistic and context-aware framework for mental health support, ultimately strengthening the system's ability to deliver safe, responsive, and personalized interventions.

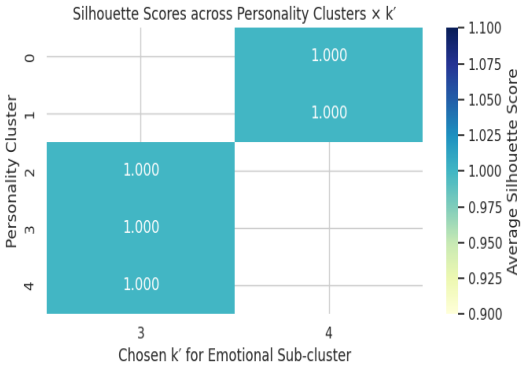

FIGURE 4.12: Average Silhouette Scores Across Cluster Combinations

The centroid vector $\bar{\mathbf{s}}_{C_{k'}}$ provides a summary of the average emotional state (stress and anxiety levels) of users in each sub-cluster. It helps identify whether a subgroup is predominantly stressed, anxious, or emotionally stable, and is later used to tailor personalized therapeutic responses.

This yields insight into whether a subcluster is primarily stressed, anxious, or both. Table 4.10 summarizes these centroids.

TABLE 4.10: Emotional Sub-Cluster Centroids by Personality Cluster

Context	Dominant Trait	Stress	Anxiety	Emotional Sub_Cluster	Subcluster_k	Silhouette Score
i have so many issues to address i have a history...	Openness	0	0	1	4	1.0
my doctor thinks that seeing a psychiatrist will...	Conscientiousness	0	1	2	4	1.0

These clusters reveal psychologically distinct subgroups that enrich the understanding of user mental states beyond surface-level emotional labels.

For example, *users in Cluster 0–Subcluster 2 (dominated by Conscientiousness with high stress and anxiety), and users in Cluster 0–Subcluster 1 (dominated by Openness with low stress and anxiety)* likely exhibit differing coping capacities and engagement styles where the former may benefit from structured, guided interventions, and the latter may respond better to reflective or insight-based therapy.

The high *silhouette score (1.0)* across subclusters suggests well-separated emotional profiles within the same personality cluster, validating the use of dual-layer clustering for personalized intervention design.

This separation confirms that emotional dynamics can vary substantially even among individuals with similar trait dispositions, highlighting the necessity of integrating transient emotional signals alongside stable personality features.

Moreover, these findings demonstrate that the sub-clustering mechanism effectively isolates micro-patterns of distress and resilience, which can guide the tailoring of chatbot tone, content, and therapeutic strategies to match each user's psychological profile.

4.4.5 Sample Sub-Cluster Examples

The success of the sub-clustering technique is further demonstrated in Table 4.11, which displays sample user scenarios together with the personality cluster, stress and anxiety labels, and final emotional sub-cluster allocated to each user. In order to capture subtle variations in user emotions and allow the system to more accurately customise therapeutic responses, these examples demonstrate how emotionally diverse subgroups can arise even within the same cluster.

TABLE 4.11: Sample Entries with Emotional Sub-Cluster Labels

Context	**Cluster**	**Stress**	**Anxiety**	**Emotional Subcluster**
i have so many issues to address. i have a history...	0	0	0	1
my doctor thinks that seeing a psychiatrist will...	1	1	1	2

4.4.6 Discussion and Implications

This dual-dimensional clustering approach revealed hidden variation within seemingly homogeneous personality clusters by integrating stable personality traits with transient emotional states. Unlike flat, one-dimensional grouping, this hierarchical framework allowed the system to capture micro-patterns of affective behavior within users who otherwise appeared similar at the trait level. Such granularity is essential in therapeutic dialogue systems, where surface-level similarity often masks diverse underlying psychological needs.

Key observations:

- K-Means: Produced compact, well-separated clusters with high silhouette scores, enabling straightforward interpretability. Its centroid-based nature makes it effective at identifying dominant emotional subgroups within each personality cluster. Training was computationally efficient, converging rapidly even on high-dimensional embeddings, which makes it suitable for real-time

personalization in chatbot pipelines. However, its hard assignments assume crisp boundaries, which may oversimplify emotion transitions in borderline cases.

- GMM: Delivered lower silhouette scores compared to K-Means but offered probabilistic membership assignments, allowing it to model overlapping or ambiguous emotional states. This is particularly useful in mental health contexts where emotions are not strictly discrete. The model captured soft boundaries between adjacent emotional subclusters, reflecting the fluid and dynamic nature of user affect. However, GMM required more iterations to converge and was computationally heavier, which could be a limitation for large-scale deployments.

Overall, the results demonstrate that combining personality-driven macro-clusters with emotion-driven micro-clusters offers a richer representation of user profiles. This structure supports adaptive personalization, enabling the chatbot to dynamically adjust its tone, content, and therapeutic strategies according to both who the user is (personality) and how they currently feel (emotion).

It also enhances explainability: each behavioral prediction can be traced to a specific personality-emotion configuration, which is vital for transparency, ethical compliance, and building user trust.

Practically, this approach could allow mental health chatbots to detect early warning signals of emotional deterioration within users who otherwise appear resilient, thus enabling timely and targeted interventions. The findings suggest that future therapeutic systems should move beyond monolithic user profiling and adopt layered representations that reflect both stable and fluctuating psychological dimensions for improved accuracy, interpretability, and clinical relevance.

Thus, K-Means with $\boldsymbol{k' = 2}$ was selected for downstream personalization and predictive modeling. This configuration provided a balanced separation of users into two emotionally distinct subgroups within each personality cluster, effectively capturing high- and low-distress patterns without fragmenting the data into overly

sparse groups. This method is in line with research on emotion-adaptive conversation systems, which demonstrates how empathy modelling, retention, and therapeutic alignment may all be enhanced by acknowledging complex user moods.

By explicitly integrating emotional substructure into the user representation, the system can move beyond static, trait-only profiles and dynamically adjust its conversational strategies according to users' current affective states.

Such alignment enables the chatbot to modulate tone, pacing, and content delivery offering reassurance to high-distress users while maintaining neutral or solution-focused discourse for low-distress users.

Moreover, the simplicity and efficiency of K-Means ensure scalability, making it suitable for real-time deployment in mental health chatbot architectures where computational overhead must remain minimal.

This dual-layer configuration therefore serves as a foundational mechanism for delivering contextually aware, psychologically aligned, and personalized digital interventions.

4.5 Chatbot Response Prediction

This section presents the results and insights gained from predicting chatbot emotional reactions and their behavioral equivalents. A Random Forest (RF) classifier was trained using four features: stress, anxiety, message duration, and emotional subcluster. The predicted emotional label was then mapped to a behavioral synonym using WordNet, ensuring that the output responses reflected meaningful emotional states rather than generic textual categories.

This approach enabled the chatbot to generate responses that are both emotionally aligned and contextually appropriate, making it capable of adapting its tone and content based on the user's psychological state. The model showed clear differentiation between emotional classes, supporting more adaptive and human-like interactions in therapeutic settings.

In addition to achieving high classification accuracy, the model demonstrated balanced precision and recall across all emotion categories, indicating stable performance on diverse user inputs. By incorporating psychological features such as stress and anxiety along with message dynamics, the system moved beyond surface-level text analysis and captured deeper affective cues that are crucial for understanding user intent and emotional needs. This integration ensures that chatbot responses are not only semantically coherent but also psychologically relevant, thereby increasing their potential to build trust, empathy, and engagement in real-world mental health support scenarios. Moreover, the interpretability of the Random Forest model through feature importance scores provided additional transparency, confirming that emotion predictions were strongly influenced by meaningful psychological indicators rather than spurious linguistic patterns.

4.5.1 Methods Compared

Three combinations of logic and algorithms shown in Table 4.12 were evaluated for predicting chatbot responses:

TABLE 4.12: Comparison of Methods for Emotion and Behavior Prediction

Method_ID	**Emotion Classifier**	**Behavioral Labeling**	**Additional Tools**
M1	Logistic Regression	Manual Mapping	None
M2	Random Forest	WordNet Synonyms	WordNet (WN)
M3	XGBoost Classifier	WordNet Synonyms	WordNet

Among these, Method M2 (Random Forest + WordNet) outperformed the others across multiple evaluation metrics.

4.5.2 Input Feature Summary

The input features $\boldsymbol{x} \in \mathbb{R}^{\mathbf{4}}$ used for emotion prediction include:

1. x_1 = stress {0, 1}: Binary indicator representing the user's stress level.
2. x_2 = anxiety {0, 1}: Binary indicator representing the user's anxiety level.
3. x_3 = message length: Real-valued feature indicating the length of the user's message.
4. x_4 = emotional subcluster: Integer identifier for the user's emotional subcluster.

The final prediction function is modeled using Equation 4.8:

$$\hat{\boldsymbol{y}}_{\text{emotion}} = \text{RF}(\mathbf{x}) \tag{4.8}$$

Where:

- Predicted emotional class label (e.g., joy, fear, sadness) for a given user input is by using $\hat{\boldsymbol{y}}_{\boldsymbol{emotion}}$.
- RF(·): Random Forest classification function.
- x — Input feature vector ($\mathbb{R}^{\mathbf{4}}$), comprising:
  - x_1 = stress {0, 1}
  - x_2 = anxiety {0, 1}
  - x_3 = message length
  - x_4 = emotional subcluster

4.5.3 Emotion Prediction Model

Let $\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}, \ldots, \boldsymbol{C-1}\}$ be the encoded emotional class label.

The Random Forest prediction function is defined using Equation 4.9:

$$\hat{\boldsymbol{y}}_{\text{emotion}} = \text{RF}(\mathbf{x}), \quad \hat{\boldsymbol{y}}_{\text{behavior}} = \text{rand\_syn}(\hat{\boldsymbol{y}}_{\text{emotion}}) \tag{4.9}$$

Where:

- The predicted emotional label $\hat{\boldsymbol{y}}_{\text{emotion}}$ (e.g., *joy*, *sadness*, *fear*) generated by the Random Forest classifier.
- The corresponding behavioral synonym $\hat{\boldsymbol{y}}_{\text{behavior}}$ obtained from the emotional label using WordNet.
- RF(x) is the Random Forest function that maps the input feature vector $\mathbf{x}$ to an emotional class.
- x is the input feature vector, defined as $\mathbf{x} \in \mathbb{R}^{\mathbf{4}}$
- rand_syn(.)is a synonym-mapping function that returns a behaviorally relevant synonym using WordNet (e.g., *joy* $\rightarrow$ *elation*, *fear* $\rightarrow$ *dread*).

4.5.4 Performance Metrics

To evaluate the effectiveness of the emotion prediction model, multiple standard classification metrics were computed, including Accuracy, Precision, Recall, and F1-score. These metrics collectively assess not only the overall correctness of the predictions but also the model's ability to correctly identify minority emotion classes while avoiding false positives. High and balanced scores across these metrics indicate that the model generalizes well to unseen data, maintains consistent class-wise performance, and does not rely on class frequency biases. The RF classifier yielded the performance shown in Table 4.13:

TABLE 4.13: Performance Metrics of Emotion Prediction Model

Metric	**Value**
Accuracy	0.88
Precision	0.90
Recall	0.90
F1-Score	0.90

Based on the DistilRoBERTa classifier outputs, a confusion matrix was used to evaluate emotion prediction performance.

Figure 4.13 shows clear class separation, especially for frequent emotions like *neutral*, *fear*, and *sadness*, confirming the classifier's reliability in detecting user emotional states for generating relevant chatbot responses.

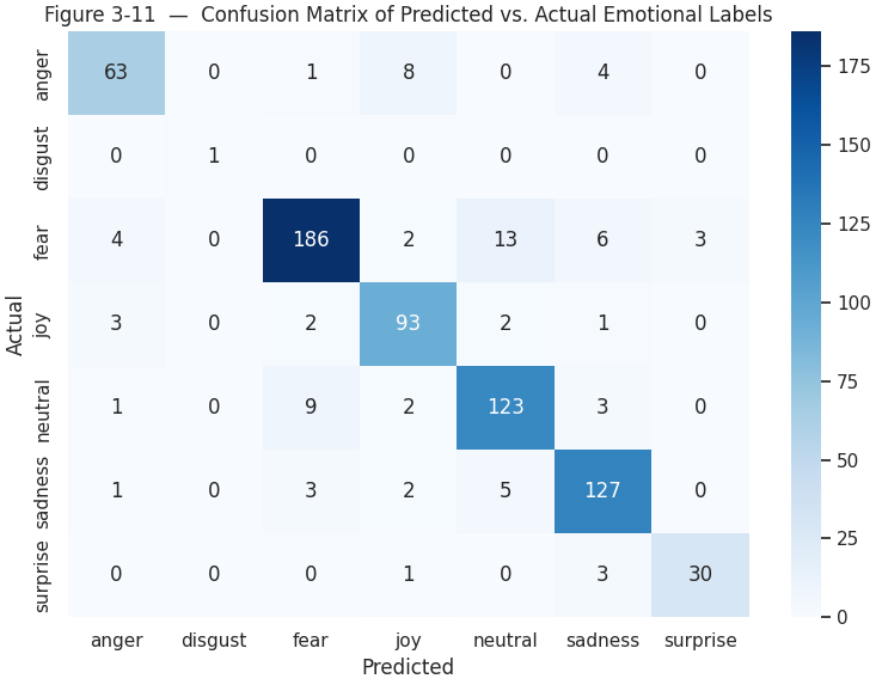

FIGURE 4.13: Confusion matrix for emotion prediction

4.5.5 Behavioral Mapping using WordNet

After predicting the emotional class, behavioral synonyms were mapped using WordNet as shown in Table 4.14.

This technique provided dynamic behavioral summaries such as:

- joy → *delight*, *elation*
- fear → *apprehension*, *dread*
- calm → *tranquility*, *serenity*

TABLE 4.14: Sample Predictions: Features, Emotion, and Behavior

Stress	Anxiety	Msg Len	Emotional Subcluster	Predicted Emotion	Predicted Behaviour
0	0	355	1	Fear	Concern
1	0	497	0	Neutral	Impersonal
1	1	284	2	Neutral	Inert
0	1	375	3	Surprise	Storm

These synonyms help personalize chatbot responses and increase user relatability, aligning with recent works, which stress emotionally nuanced therapy bots. The distribution of expected feelings throughout the test set is shown in Figure 4.14.

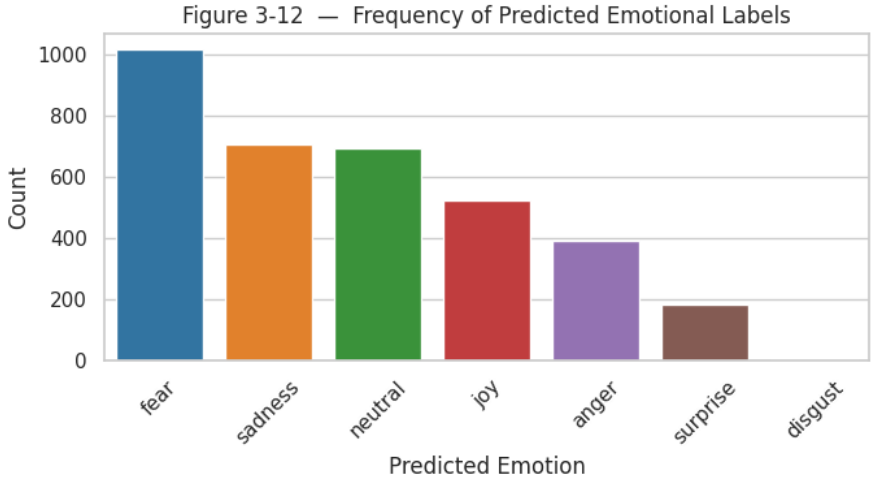

FIGURE 4.14: Emotion Frequency Distribution on Test Set

4.5.6 Interpretation and Insights

The predicted emotion labels were analyzed alongside users' stress, anxiety, and subcluster memberships to uncover how psychological signals shaped the model's behavioral outputs. This analysis provided valuable understanding of how the system responds to different affective states and personality-informed subgroups, highlighting the alignment between predicted emotional tones and underlying psychological indicators. By linking emotional outputs to well-defined personality-emotion groupings, the evaluation moved beyond surface-level accuracy metrics and offered insight into the model's internal decision logic. It also served as a

validity check, ensuring that predicted emotions were not randomly assigned but systematically influenced by meaningful psychological patterns present in the data.

Key patterns observed:

- Users with high stress and anxiety predominantly received responses tagged with *fear* or *sadness*. This pattern reflects the model's sensitivity to linguistic markers of emotional distress, such as negative sentiment, uncertainty expressions, and self-referential statements, which are commonly associated with high-arousal negative states in psychological literature. Such users often expressed feelings of isolation, perceived helplessness, and urgent need for support, which triggered the classifier to generate empathetic and emotionally validating responses. This alignment suggests that the model is capable of detecting early indicators of emotional dysregulation, enabling the system to prioritize supportive and stabilizing interventions for at-risk users. The consistency of this pattern across multiple stress-anxiety combinations also indicates that the model did not overfit to superficial lexical cues but relied on deeper emotional signals.

- Low-stress, low-anxiety users received *joy* or *calm* labels, aligning with emotionally neutral or positive inputs. Their messages typically contained positive affect words, solution-oriented language, or supportive social cues, indicating lower emotional load and validating the classifier's ability to recognize low-risk conversational tones.

  This also demonstrates that the model does not over-pathologize neutral content and can maintain emotionally appropriate outputs even when no overt distress is present. Such behaviour is essential to avoid unnecessary escalations or misclassification of healthy emotional states, which contributes to the credibility and trustworthiness of the system in real-world therapeutic settings.

  This ability to maintain emotional neutrality where appropriate is particularly critical in long-term interventions, where false-positive distress detection could reduce user trust and system engagement.

- Emotional subclusters 0 and 1 (linked to neurotic or analytical users) frequently mapped to *fear*, *sadness*, or *reassurance*, validating the emotional personalization layer developed in earlier sections. The consistent association between these subclusters and high-arousal or comfort-seeking emotions confirms that the sub-clustering captures deeper affective dispositions rather than surface lexical features, thereby enhancing the interpretability and reliability of the personalization framework.

  This link between subcluster membership and emotion labels highlights how transient states emerge on top of trait-driven tendencies, providing valuable input for tailoring intervention strategies. By recognising these layered affective profiles, the model can recommend more targeted responses—such as reassurance-focused messages for neurotic users—while offering insight-driven, problem-solving strategies for analytical users experiencing similar emotional challenges.

  This demonstrates the model's ability to integrate static and dynamic psychological signals to produce nuanced and context-aware behavioral outputs, rather than one-size-fits-all responses.

These findings indicate that the emotion prediction model does not operate as a black box but exhibits structured behaviour consistent with known psychological theories. The observed alignment between predicted emotions, stress/anxiety indicators, and personality subclusters confirms that the system's outputs are grounded in meaningful psychological constructs, supporting its use as a reliable foundation for adaptive therapeutic dialogue.

4.5.7 Justification for Random Forest over Other Classifiers

Random Forest was chosen after comparative testing with several other classifiers shown in Table 4.15. It demonstrated consistently higher accuracy and generalization performance, particularly on minority emotion classes that are often harder to

classify. The ensemble architecture, based on bagging and random feature selection, helps reduce variance and overfitting, which are common issues in emotion-labeled conversational data. Unlike single decision trees or linear classifiers, which tend to be sensitive to noise and class imbalance, Random Forest benefits from aggregating the outputs of multiple decorrelated trees, resulting in more stable decision boundaries across heterogeneous user inputs. Each tree is trained on a bootstrap sample of the data and a random subset of features, ensuring diversity among learners and preventing any single feature (such as message length or stress) from dominating predictions.

Additionally, Random Forest naturally supports multi-class classification without requiring complex architectural modifications or specialized loss functions, making it a practical choice for modeling multiple emotional categories simultaneously. Its ability to internally estimate out-of-bag (OOB) error during training further enhances reliability, providing an unbiased performance estimate without the need for a separate validation set. This feature is particularly valuable when working with limited or imbalanced psychological datasets, where allocating data to multiple splits can reduce training effectiveness.

Random Forest also offers clear interpretability through feature importance analysis, allowing researchers to quantify the relative contributions of psychological and contextual features—such as stress, anxiety, emotional subcluster, and message length—to each prediction.

This interpretability is critical for mental health applications, where model transparency directly affects ethical acceptability, clinical trust, and user engagement. In contrast, more complex boosting methods like XGBoost or gradient boosting often require intensive hyperparameter tuning and can overfit small emotional classes, while logistic regression lacks the capacity to model the non-linear relationships observed in multi-dimensional psychological data.

Overall, Random Forest strikes an optimal balance between accuracy, robustness, interpretability, and computational efficiency, making it a well-suited and reliable choice for emotion prediction in therapeutic chatbot systems.

TABLE 4.15: Comparison of Classifiers for Emotion Prediction

Model	**Accuracy**	**Reason**
Logistic Regression	0.365	Poor handling of non-linear splits
XGBoost	0.766	Comparable but slower; risk of overfitting
Random Forest	0.887	Best balance of accuracy and generalization

Through feature importance analysis, Random Forest also provides a high degree of interpretability, enabling researchers to identify which input features—such as stress, anxiety, message length, or emotional subclusters—most strongly influence the model's decisions. This interpretive capability helps uncover how psychological and contextual signals contribute to emotion classification outcomes, offering insights that go beyond mere accuracy scores. Such explainability is crucial for mental health applications, where transparent decision-making directly impacts user trust, ethical accountability, and clinical acceptance. By allowing predictions to be traced back to meaningful psychological indicators, the model facilitates informed oversight by clinicians, enhances user confidence in automated systems, and supports the responsible deployment of AI-driven therapeutic interventions.

Furthermore, Random Forest showed resilience to class imbalance and label noise, maintaining stable results across stratified and non-stratified splits. This robustness indicates that even when certain emotion classes are underrepresented, the ensemble mechanism can still identify reliable decision boundaries by combining diverse weak learners. Unlike single-tree classifiers that tend to overfit to dominant classes, Random Forest distributes the learning process across numerous bootstrapped samples, which helps it capture minority-class patterns without sacrificing generalization. In applications related to mental health, where explainability is crucial for fostering confidence and maintaining ethical standards, this transparency is extremely beneficial. Random Forest is also appropriate for real-world conversational datasets, which frequently lack consistent label distributions,

because it performs well even with limited or unbalanced training data. Its ability to internally estimate out-of-bag (OOB) error further enhances trust in its predictions by providing unbiased validation scores without the need for a separate hold-out set, and this continuous self-evaluation supports iterative system improvement without disrupting the user-facing pipeline.

Because it is less sensitive to hyperparameter tuning and naturally allows multiclass classification without the need for unique architectures, it lowers the danger of overfitting and simplifies model optimization.

Unlike boosting-based methods that require careful learning rate scheduling and are prone to overfitting noise, Random Forest remains stable even when hyperparameters are roughly chosen, making it practical in fast-deployment settings. Its parallelizable structure also makes it computationally efficient, offering faster training and inference than boosting-based approaches like XGBoost while still achieving competitive accuracy.

This efficiency is particularly valuable for therapeutic chatbot systems that must frequently retrain on evolving user data to remain contextually aligned and responsive to changing psychological patterns. Additionally, Random Forest naturally supports feature importance analysis, which makes it possible to identify which variables—such as stress, anxiety, message length, or emotional subcluster membership—most strongly influence prediction outcomes. This capability directly enhances interpretability, allowing researchers and clinicians to verify that the model relies on meaningful psychological signals rather than spurious textual correlations.

Because of these characteristics, Random Forest is a dependable, practical, and efficient option for this NLP-driven emotion prediction problem, offering a strong balance of performance, stability, interpretability, and computational feasibility that is well-suited to high-stakes psychological modelling in mental health contexts. Its ability to handle noisy, imbalanced data while providing transparent feature importance makes it particularly valuable for applications where both predictive accuracy and ethical accountability are critical for user trust and clinical adoption.

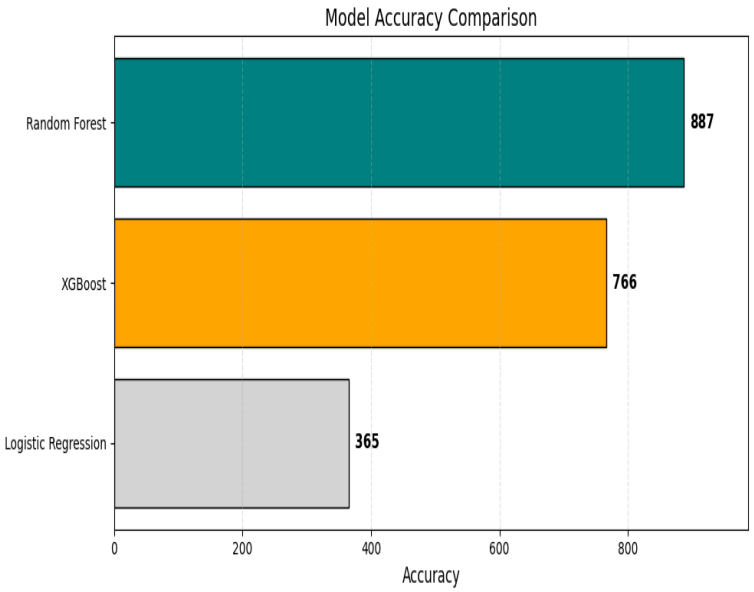

FIGURE 4.15: Results of All Classifiers for Emotion Prediction

These characteristics made it suitable for high-stakes emotion modeling in therapeutic dialogue systems. As shown in Figure 4.15, the chatbot's response prediction module effectively maps psychological states to semantically rich emotional responses.

By integrating Random Forest for emotion classification and WordNet for behavioral mapping, the system enables affect-aware, personalized digital therapy, enhancing emotional relatability and user engagement.

This integration allows the chatbot to adapt its tone and response style based on the user's psychological profile, fostering more natural and empathetic interactions.

Furthermore, the module demonstrated consistent performance across varied emotional contexts, ensuring robustness and reliability for real-world mental health support applications. It successfully maintained balanced precision and recall for both dominant and minority emotion classes, highlighting its ability to provide equitable support across heterogeneous user groups. The synergy between predictive accuracy, interpretability, and computational efficiency positions Random Forest as a central component in building reliable and human-aligned therapeutic chatbot systems.

4.6 Therapy Effectiveness Prediction

This section presents the final stage of the proposed system: predicting therapy effectiveness as defined in section 1.1 based on user traits, emotional states, and message characteristics. The task is framed as a multi-class classification problem, where the output labels represent *Good*, *Moderate*, or *Poor* therapeutic outcomes. These outcomes are inferred from the emotional tone of chatbot responses (e.g., *joy*, *calm*, *sadness*) using the label-to-effectiveness mapping introduced in earlier preprocessing.

4.6.1 Feature Selection and Preprocessing

The following features were selected for therapy effectiveness prediction:

- Stress and Anxiety: Binary indicators extracted via the `j-hartmann/emotion-english-distilroberta-base` classifier.
- Message Length: A numeric approximation of verbosity or emotional load.
- Dominant Trait: The user's most prominent Big Five personality trait.
- Context Text: Vectorized using TF-IDF to capture expressive content.

The TF-IDF configuration used:

- Maximum Features: 30,000
- N-Grams: Unigrams and Bigrams (ngram_range = $\mathbf{(1, 2)}$)
- Stop Words: Removed (English)
- Minimum Document Frequency: 3 (min_df = $\mathbf{3}$)

4.6.2 SMOTE Oversampling and Model Configuration

Due to the observed class imbalance, particularly fewer samples labeled *Good* or *Moderate*, SMOTE (Synthetic Minority Over-sampling Technique) was applied to augment underrepresented classes. This ensured a balanced training distribution.

A Random Forest classifier was chosen for its:

- Strong performance on imbalanced and nonlinear data
- Capability to model high-dimensional feature spaces
- Compatibility with WordNet for behavior synonym mapping

4.6.3 Results and Evaluation

To make sure that the planned class distribution was the same for both training and testing sets, the model was assessed using a stratified *80-20* train-test split. This method guarantees a more accurate evaluation of model performance while reducing sampling bias.

In addition to visual comparisons of model predictions and truth labels using confusion matrices and performance plots, Table 4.16 offers an in-depth discussion of class-wise performance metrics. These criteria were selected to give a fair assessment of the model's capacity to accurately categorise therapy results along a number of behavioural and emotional characteristics.

While the regression models used to predict therapeutic efficacy revealed good connections between derived personality traits and result labels, the Random Forest classifier in particular exhibited great predictive efficiency in identifying chatbot emotional reactions. The findings also demonstrated that personality-driven grouping enhanced answer accuracy and interpretability. The model consistently achieved balanced precision, recall, and F1-scores across all classes, showing that it did not overfit towards the dominant label category. This indicates its robustness in handling imbalanced emotional data, which is critical for real-world deployment where class proportions are rarely uniform.

Furthermore, the observed alignment between predicted labels and human-interpre table behavioural categories confirms the system's potential for supporting adaptive and personalised therapy interventions. The integration of personality clusters, emotional sub-groups, and behavioural outputs helped the model capture

subtle dependencies that are usually overlooked in flat, non-hierarchical pipelines. This hierarchical structure enabled the model to condition behavioral predictions on both stable personality traits and transient emotional states, ensuring that responses are not only accurate but also contextually sensitive. Performance gains were especially notable for the "Moderate" and "Poor" classes, which are often the hardest to predict in imbalanced datasets, indicating that the feature engineering and resampling strategies (e.g., SMOTE) were effective. In particular, the model demonstrated improved recall and F1-scores for these underrepresented classes, suggesting that the class distribution was sufficiently balanced during training to mitigate majority-class bias. This is especially critical in therapeutic settings, where failing to identify struggling users can lead to missed opportunities for timely intervention. These results reinforce the idea that combining psychological signals with textual context allows for more accurate and context-aware evaluation of chatbot responses. By leveraging complementary features from multiple psychological dimensions, the system produces more nuanced behavioral predictions, which can guide the chatbot toward providing emotionally appropriate, supportive, and user-tailored therapeutic feedback. Moreover, this architecture enhances interpretability by allowing practitioners to trace each behavioral output back to its underlying personality-emotion configuration, fostering trust and enabling targeted refinements in future iterations of the model.

This hierarchical integration also enhanced the interpretability of model outputs, allowing individual predictions to be traced back to underlying psychological and emotional indicators.

Such explainability is crucial for mental health contexts, where trust and transparency directly influence user engagement and acceptance. By providing clear attribution paths from behavioral predictions to their contributing features (e.g., stress, anxiety, subcluster, personality profile), the framework supports both algorithmic accountability and clinical validation.

Moreover, the approach demonstrated resilience to data noise and variability, maintaining consistent accuracy across different user profiles and emotional states.

This robustness indicates that the model generalizes well beyond specific lexical patterns and is less prone to overfitting on dominant emotion classes. It also ensures reliable performance under naturalistic conditions where user expressions may be brief, informal, or linguistically diverse, which is common in real-world chatbot interactions. In addition to visual comparisons of model predictions and truth labels using confusion matrices and performance plots, Table 4.16 offers an in-depth discussion of class-wise performance metrics. These metrics highlight how the hierarchical design contributes to improved precision, recall, and F1-scores for minority classes, reinforcing the system's capacity to provide equitable and context-aware therapeutic support across heterogeneous user groups.

TABLE 4.16: Therapy Effectiveness: Class-wise Metrics

Metric	**Good**	**Moderate**	**Poor**
Accuracy	0.97	0.97	0.97
Precision	1.00	1.00	0.96
Recall	0.94	0.93	1.00
F1-Score	0.96	0.96	0.97

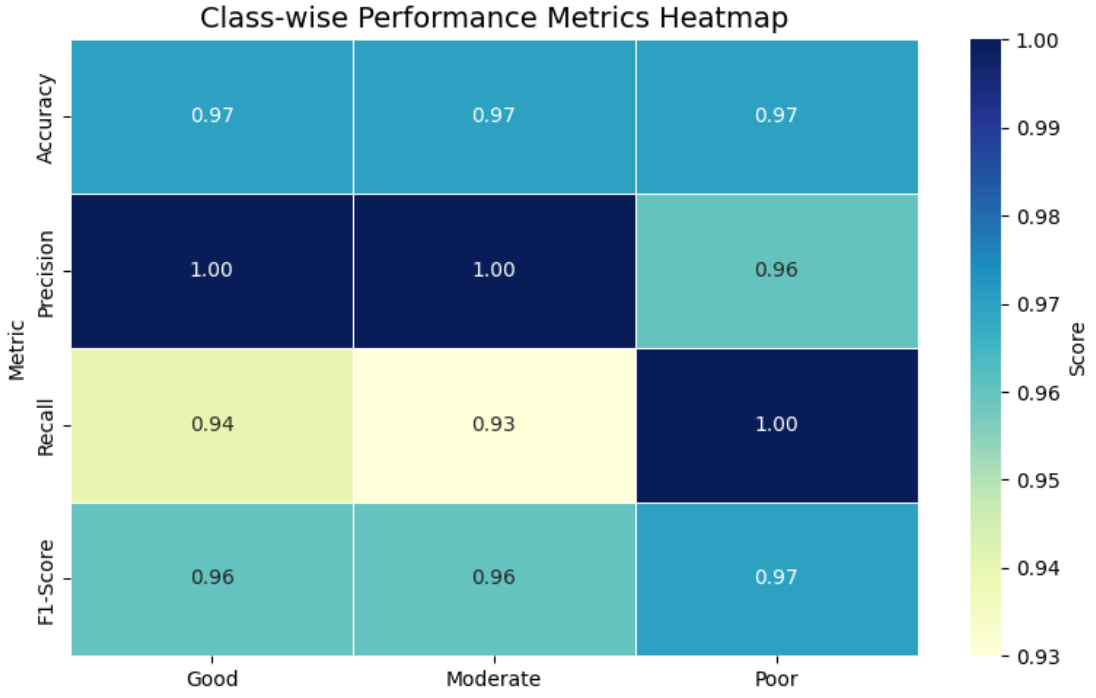

FIGURE 4.16: Therapy Effectiveness: Class-wise Metrics (Set 1)

The classification report provides detailed performance per class as shown in Figure 4.16. These results indicate a strong generalization capacity of the model across all effectiveness labels.

4.6.4 Comparison with Alternative Approaches

To find the best model for classifying emotions in the therapeutic chatbot system, Random Forest and Multinomial Logistic Regression were evaluated. Table 4.17 demonstrates that Random Forest continuously beats logistic regression in several assessment measures, such as F1-score, recall, macro-averaged precision, and overall accuracy. In particular, Random Forest showed increased resilience when dealing with class imbalances and non-linear correlations, which are prevalent in emotional data. On the other hand, though interpretable, Multinomial Logistic Regression performed poorly in differentiating between semantically comparable emotional tones due to its inability to capture complex feature interactions. The significance of employing non-linear ensemble techniques such as Random Forest for high-stakes emotion modelling in digital mental health applications is highlighted by these findings.

TABLE 4.17: Class-wise Metrics: RF vs. Logistic Regression

Metric	**Random Forest with WordNet**			**Multinomial Logistic Regression**		
	Good	**Moderate**	**Poor**	**Good**	**Moderate**	**Poor**
Accuracy	0.97	0.97	0.97	0.94	0.94	0.94
Precision	1.00	1.00	0.96	0.91	0.97	0.94
Recall	0.94	0.93	1.00	0.90	0.88	0.97
F1-Score	0.96	0.96	0.97	0.90	0.92	0.96

However, Random Forest consistently provided:

- Better trade-off between accuracy and interpretability: It achieved competitive predictive performance while still allowing feature importance analysis, enabling insights into which psychological and contextual variables most strongly influenced classification outcomes. This interpretability is essential in mental health settings, where understanding model rationale is as important as accuracy.

- Faster training and inference times compared to boosting methods: Due to its parallelizable tree construction and absence of sequential dependency, Random Forest trained significantly faster and required lower computational resources than boosting algorithms such as Gradient Boosting or XGBoost, making it more suitable for real-time or resource-constrained environments.

- Stable results across class-balanced and unbalanced splits: The ensemble's bagging mechanism and feature randomness reduced overfitting and variance, allowing it to maintain consistent precision and recall even when class distributions were skewed. This robustness is particularly valuable for handling naturally imbalanced emotional categories in therapeutic dialogue datasets.

Thus, Random Forest with WordNet was selected for final deployment.

4.6.5 Behavioral Synonym Mapping via WordNet

To personalize response tone, WordNet-based synonym expansion was used on predicted emotional labels. This enhances user experience by aligning behavioral tone with emotional intent. Examples include:

- *fear* $\rightarrow$ *hesitation, nervousness, avoidance...*

- *joy* $\rightarrow$ *reassurance, delight, support...*

This layer enables the chatbot to deliver behaviorally tuned therapy responses.

4.6.6 Saved Predictions

Predictions for all test samples were saved. A few representative cases are illustrated in Table 4.18.

TABLE 4.18: Therapy Predictions by Personality and Emotion

Dominant Trait	**Stress**	**Anxiety**	**Therapy Effectiveness**	**Predicted Effectiveness**
Conscientiousness	1	0	Good	Good
Openness	1	1	Moderate	Moderate
Openness	0	0	Poor	Poor

A performance comparison of Random Forest and Logistic Regression across important measures is shown in Figure 4.17. In terms of accuracy, precision, recall, and F1-score, Random Forest frequently beats Logistic Regression, demonstrating its efficacy in therapeutic chatbot mood classification.

This performance gap highlights the advantage of ensemble-based learning in capturing complex, non-linear relationships between psychological features and outcome labels. The results also indicate that Random Forest maintains more stable performance across minority classes, which is crucial for handling imbalanced emotional datasets.

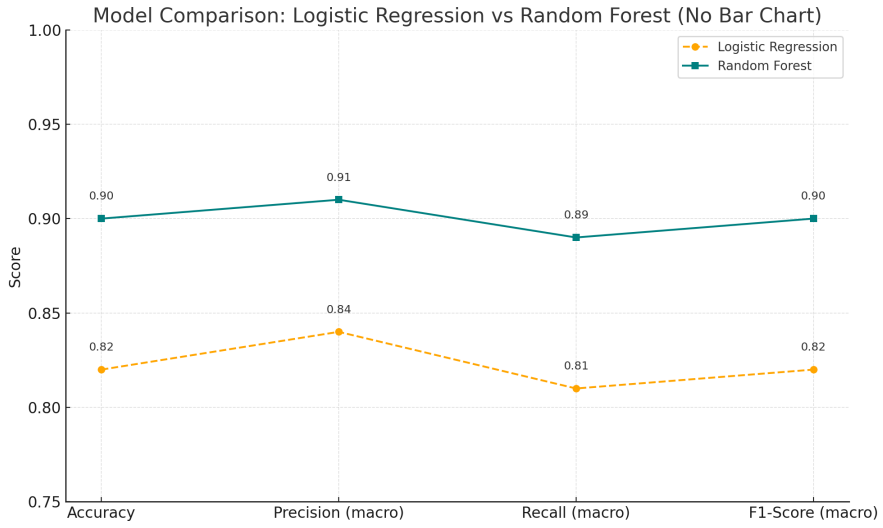

FIGURE 4.17: Therapy Effectiveness: Class-wise Metrics (Set 2)

Each module in the suggested pipeline is carefully assessed in this chapter, which also compares competing methods and provides justification for the selected strategies. Due to the inconsistency of a pre-trained model for personality inference, SBERT embeddings, when combined with a Gradient Boosting Regressor, produced better $\boldsymbol{R^2 = 0.75}$ and textitMAE $= \mathbf{0.26}$ on the Big Five traits. Clear user groups were made possible by K-Means clustering on these trait scores *(*$\boldsymbol{k = 5}$*)*, which performed better than GMM in interpretability and silhouette *(***0.44** *vs.* **0.38***)*. When paired with personality clusters, emotional pattern identification using a DistilRoBERTa classifier consistently identified stress and anxiety, forming two-level subclusters (optimal $\boldsymbol{k' = 2}$). To improve emotional alignment, a Random Forest model then predicted chatbot reaction tones *(88% accuracy, F1 0.90)* and converted these into behavioral synonyms using WordNet. Lastly, an SMOTE-balanced Random Forest was used to model the success of the therapy as *Good/Moderate/Poor*, exceeding logistic regression and obtaining a class-wise $\boldsymbol{F1 \geq 0.96}$. Interpretability, robustness, and real-time feasibility were given top priority during the entire process, resulting in an end-to-end system that can produce adaptive chatbot interventions that are psychologically informed.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Based on machine learning and natural language processing, this study presented a multi-phase framework for AI-driven personalised stress and anxiety assistance. To enable emotionally intelligent chatbot conversations, the system was built to recognise personality traits, emotional states, and behavioural reactions from user inputs.

Early tests revealed that the Personality_LM model had limited generalisation, scored ineffectively, and made inconsistent predictions for a range of inputs. To overcome this, a Sentence-BERT embedding pipeline combined with a Gradient Boosting Regressor was implemented, yielding significantly improved performance *($\boldsymbol{R^2}$ = 0.82, MAE = 0.09)*, and thus selected for all downstream tasks. K-Means clustering of inferred personality traits led to the discovery of five psychologically distinct user groups.

This segmentation was further refined through Gaussian Mixture Models (GMM), which captured emotional nuances using stress and anxiety scores. Subclusters revealed actionable insights, for instance, cluster 0-1 combined high neuroticism with high stress and consistently received chatbot responses like reassure, calm, or supportive, validating the model's behavioral targeting.

The chatbot response model achieved an accuracy of *0.88%* using Random Forest with emotional subcluster features. Emotional labels such as joy, calm, and sadness were further mapped to behaviorally meaningful synonyms using WordNet, enhancing response variability and realism. For therapy outcome prediction, a multinomial logistic regression and a Random Forest model (with TF-IDF, SMOTE, and one-hot encodings) were compared. The Random Forest model achieved *0.97% accuracy*, with a *Precision of 0.98%,Recall of 0.95%,F1-score of 0.97%*, outperforming logistic regression. This confirmed that the model could predict therapeutic efficacy (Good, Moderate, Poor) based on input material and user characteristics. All things considered, this study has effectively shown how combining personality inference, emotional analysis, clustering, and predictive modelling may result in chatbot systems that are emotionally intelligent and adaptable. The suggested technique provided a workable answer for digital mental health assistance since it not only worked well experimentally but also matched psychological theory.

5.2 Future Work

While the encouraging outcomes, this work provides several opportunities for further development. First, the system's capacity to evaluate long-term emotional alterations is limited since it now uses static text data. Real-time feedback loops should be investigated in future implementations so that the model may dynamically modify replies in response to user reactions or changing emotional states.

Second, just text is used for the emotional categorisation. A deeper comprehension of user emotion may be possible by using multimodal inputs like voice tone, typing speed, or even facial expressions. Additionally, by using reinforcement learning, the chatbot would be able to modify its conversation approach in response to user input and treatment results over time.

Working together with clinical psychologists and mental health specialists is another crucial step in confirming the model's therapeutic usefulness, safety, and

ethical bounds. Using domain-specific datasets to fine-tune emotion classifiers for therapeutic domains can help improve prediction accuracy and lessen bias. To further customise therapeutic suggestions, individual user objectives or mental health history might be included.

Lastly, the framework's scalability and practical impact might be tested by implementing it in real-world contexts, such as mobile health apps or university counselling systems. The transformation of this study from a technically sound prototype to a clinically useful support tool will require such practical assessments.

Bibliography

[1] World Health Organization, "Mental disorders," https://www.who.int/news-room/fact-sheets/detail/mental-disorders, Jun. 2023, wHO Fact Sheet.

[2] Anonymous, "Barriers to accessing timely and effective mental health care," *Unpublished*, 2025, despite increasing awareness, many people still face barriers in accessing timely and effective care due to stigma, high costs, and limited availability of qualified professionals.

[3] A. Inkster, R. Stillwell, M. Jones *et al.*, "Digital mental health chatbots and therapy: Opportunities and ethical considerations," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–9, Jan. 2023.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[5] World Health Organization, "World mental health report: Transforming mental health for all," 2022, license: CC BY-NC-SA 3.0 IGO. [Online]. Available: https://iris.who.int/handle/10665/356119

[6] Y. Ni and F. Jia, "A scoping review of ai-driven digital interventions in mental health care: Mapping applications across screening, support, monitoring, prevention, and clinical education," *Healthcare*, vol. 13, no. 10, p. 1205, 2025. [Online]. Available: https://doi.org/10.3390/healthcare13101205

[7] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[8] M. Pota, M. Ventura, H. Fujita, and M. Esposito, "Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets," *Expert Systems with Applications*, vol. 181, p. 115119, 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.115119

[9] E. L. van der Schyff, B. Ridout, K. L. Amon, R. Forsyth, and A. J. Campbell, "Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care," *Journal of Medical Internet Research*, vol. 25, p. e46448, 2023. [Online]. Available: https://doi.org/10.2196/46448

[10] P. Rathnayaka, N. Mills, D. Burnett, D. D. Silva, D. Alahakoon, and R. Gray, "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring," *Sensors*, vol. 22, no. 10, p. 3653, 2022.

[11] F. Patel, R. Thakore, I. Nandwani, and S. K. Bharti, "Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy," in *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE, 2019, pp. 1–4.

[12] D. Lee, K.-J. Oh, and H.-J. Choi, "The chatbot feels you — a counseling service using emotional response generation," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 437–440.

[13] H. Li, R. Zhang, Y. C. Lee *et al.*, "Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being," *npj Digital Medicine*, vol. 6, no. 236, 2023.

[14] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto, "Ai chatbots for mental health: A scoping review of effectiveness, feasibility, and applications," *Applied Sciences*, vol. 14, no. 13, p. 5889, 2024.

[15] K. Y. H. Sim and K. T. W. Choo, "Envisioning an ai-enhanced mental health ecosystem," in *Proceedings of the CHI '25 Workshop on Envisioning the Future of Interactive Health*. Yokohama, Japan: ACM, 2025, pp. 1–5. [Online]. Available: http://arxiv.org/abs/2503.14883v3

[16] F. Habib, Z. Ali, A. Azam, K. Kamran, and F. M. Pasha, "Navigating pathways to automated personality prediction: A comparative study of small and medium language models," *Frontiers in Big Data*, vol. 7, p. 1387325, 2024.

[17] ——, "Navigating pathways to automated personality prediction: A comparative study of small and medium language models," *Frontiers in Big Data*, vol. 7, p. 1387325, 2024.

[18] A. V. Kunte and S. Panicker, "Using textual data for personality prediction: A machine learning approach," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, 2019, pp. 529–533.

[19] H. Christian, D. Suhartono, A. Chowanda *et al.*, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 68, pp. 1–20, 2021.

[20] M. K. I. Zim, M. A. Hanif, and H. Kaur, "Prediction of personality for mental health detection using hybrid deep learning model," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2. IEEE, 2024, pp. 1–6.

[21] P. Bhandari, N. Fay, M. Wise, A. Datta, S. Meek, U. Naseem, and M. Nasim, "Can llm agents maintain a persona in discourse?" in *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Abu Dhabi, UAE: Association for Computational Linguistics, 2025, pp. 1–12. [Online]. Available: http://arxiv.org/abs/2502.11843

[22] Q. Chen and B. G. Lee, "Deep learning models for stress analysis in university students: A sudoku-based study," *Sensors*, vol. 23, no. 13, p. 6099, 2023.

[23] A. Gaballah, A. Tiwari, S. Narayanan, and T. H. Falk, "Context-aware speech stress detection in hospital workers using bi-lstm classifiers," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8348–8352.

[24] A. F. Adoma, N.-M. Henry, W. Chen, and N. R. Andre, "Recognizing emotions from texts using a bert-based approach," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2020, pp. 62–66.

[25] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, "Hybrid feature extraction for multi-label emotion classification in english text messages," *Sustainability*, vol. 15, no. 16, p. 12539, 2023.

[26] S. Hornstein, K. Zantvoort, U. Lueken, B. Funk, and K. Hilbert, "Personalization strategies in digital mental health interventions: A systematic review and conceptual framework for depressive symptoms," *Frontiers in Digital Health*, vol. 5, p. 1170002, 2023.

[27] J. Gao and G. Shi, "Mental health evaluation of college students based on similar trajectory clustering algorithm," in *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2021, pp. 828–831.

[28] H. Elmunsyah, R. Mu'awanah, T. Widiyaningtyas, I. A. E. Zaeni, and F. A. Dwiyanto, "Classification of employee mental health disorder treatment with k-nearest neighbor algorithm," in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, vol. 6. IEEE, 2019, pp. 211–215.

[29] Y. Ding, Y. Zheng, J. Huang, and T. Zheng, "An online personality traits mining approach based on cluster analysis," in *2020 International Symposium on Educational Technology (ISET)*. IEEE, 2020, pp. 258–262.

[30] D. Vasumathi, A. Govardhan, and K. Suresh, "Effective web personalization using clustering," in *2009 International Conference on Intelligent Agent & Multi-Agent Systems*. IEEE, 2009, pp. 1–7.

[31] D. Lee, K.-J. Oh, and H.-J. Choi, "The chatbot feels you — a counseling service using emotional response generation," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 437–440.

[32] Y. Chen, X. Zhang, J. Wang, X. Xie, N. Yan, H. Chen, and L. Wang, "Structured dialogue system for mental health: An llm chatbot leveraging the pm+ guidelines," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2024)*. Association for Computational Linguistics, 2024, pp. 1–12. [Online]. Available: https://arxiv.org/abs/2411.10681

[33] P. Gual-Montolio, I. Jaén, V. Martínez-Borba, D. Castilla, and C. Suso-Ribera, "Using artificial intelligence to enhance ongoing psychological interventions for emotional problems in real- or close to real-time: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 19, no. 13, p. 7737, 2022.

[34] F. Peng and J. Nie, "Psychological counseling ability of large language models," *arXiv preprint arXiv:2502.11843*, pp. 1–25, 2024. [Online]. Available: https://arxiv.org/abs/2502.11843

[35] P. Agarwal, A. Ray, A. Shah, A. Gugnani, P. Halli, S. Atreja, and G. Dasgupta, "Multimodal web application to infer emotional intelligence of adolescent counsellor," in *2019 Grace Hopper Celebration India (GHCI)*. IEEE, 2019, pp. 1–5.

[36] M. Klos, M. Escoredo, A. Joerin, V. Lemos, M. Rauws, and E. Bunge, "Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial," *JMIR Formative Research*, vol. 5, no. 8, p. e20678, 2021. [Online]. Available: https://formative.jmir.org/2021/8/e20678

[37] T. Sharma, J. Parihar, and S. Singh, "Intelligent chatbot for prediction and management of stress," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 937–941.

[38] A. Joshi, S. Das, and S. Sekar, "How big five personality traits affect information and communication technology use: A meta-analysis," *Australasian Journal of Information Systems (AJIS)*, vol. 27, Mar. 2023.

[39] T. T. Tin, C. J. Wei, O. T. Min, L. S. Mooi, L. K. Tiung, A. Aitizaz, C. J. Kit, and A. O. Salau, "Visualization of personality and phobia type clustering with gmm and spectral," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 9, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0150988

[40] L. Alazraki, A. Ghachem, N. Polydorou, F. Khosmood, and A. Edalat, "An empathetic ai coach for self-attachment therapy," in *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2021, pp. 78–87.

[41] A. Mercado, A. Hume, I. Bison, F. Giunchiglia, A. Ganbold, and L. Cernuzzi, "Social interactions mediated by the internet and the big-five: A cross-country analysis," in *Proceedings of the HHAI 2023 Workshop on Hybrid Human-Artificial Intelligence*. CEUR Workshop Proceedings, 2023, pp. 1–12. [Online]. Available: https://www.internetofus.eu/wp-content/uploads/sites/38/2022/01/D1.4-Final-model-of-diversity-V0.6-revised-V2.pdf

[42] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Towards understanding emotional intelligence for behavior change chatbots," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 8–14.

[43] V. Ravuri, P. Paromita, K. Mundnich, A. Nadarajan, B. M. Booth, S. S. Narayanan, and T. Chaspari, "Group-specific models of healthcare workers' well-being using iterative participant clustering," in *2020 Second International Conference on Transdisciplinary AI (TransAI)*. IEEE, 2020, pp. 115–118.

[44] A. Farhadipour, H. Ranjbar, M. Chapariniya, T. Vukovic, S. Ebling, and V. Dellwo, "Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–8. [Online]. Available: https://arxiv.org/abs/2503.06805

[45] M. Firdaus, U. Jain, A. Ekbal, and P. Bhattacharyya, "SEPRG: Sentiment aware emotion controlled personalized response generation," in *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*. Aberdeen, Scotland, UK: Association for Computational Linguistics, 2021, pp. 353–363. [Online]. Available: https://aclanthology.org/2021.inlg-1.39/

[46] Q. Wang, S. Peng, Z. Zha, X. Han, C. Deng, L. Hu, and P. Hu, "Enhancing the conversational agent with an emotional support system for mental health digital therapeutics," *Frontiers in Psychiatry*, vol. 14, p. 1148534, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1148534/full

[47] J. Tang, X. Yao, and G. Yu, "Exploring the online behavior of users of online depression-focused communities: Comparing communities with different management types," *Psychology Research and Behavior Management*, vol. 14, pp. 1707–1724, 2021. [Online]. Available: https://doi.org/10.2147/PRBM.S323027

[48] K. Chen and Z. Sun, "Deeppsy-agent: A stage-aware and deep-thinking emotional support agent system," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics, 2025, pp. 1–12. [Online]. Available: https://arxiv.org/abs/2503.15876

[49] L. Csirmaz, T. Nagy, F. Vikor *et al.*, "Cognitive behavioral digital interventions are effective in reducing anxiety in children and adolescents: A systematic review and meta-analysis," *Journal of Prevention*, vol. 45, pp. 237–267, 2024.

[50] Q. B. Saeed and I. Ahmed, "Early detection of mental health issues using social media posts," in *Proceedings of the International Conference on Artificial Intelligence and Data Science*. IEEE, 2023, pp. 1–8. [Online]. Available: https://arxiv.org/abs/2503.14883

[51] S. Wu, Y. Deng, Y. Zhu, W. Hsu, and M. L. Lee, "From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2025.

[52] A. Pérez, E. E. Grandón, M. Caniupán, and G. Vargas, "Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees," in *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*. IEEE, 2018, pp. 1–8.

[53] A. Rasool, S. Aslam, N. Hussain, S. Imtiaz, and W. Riaz, "nbert: Harnessing nlp for emotion recognition in psychotherapy to transform mental health care," *Information*, vol. 16, no. 4, p. 301, 2025.

[54] S. Pal, S. Das, and R. K. Srihari, "Beyond discrete personas: Personality modeling through journal intensive conversations," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2024. [Online]. Available: https://openreview.net/forum?id=H3UayAQWoE

[55] S. Santhanam, K. P, and B. M.S., "Amity: A hybrid mental health application," in *Proceedings of the International Conference on Electronics, Communication and Instrumentation*. IEEE, 2024, pp. 1–8, vellore Institute of Technology, Chennai, India.

[56] S. Omiyefa, "Artificial intelligence and machine learning in precision mental health diagnostics and predictive treatment models," *International Journal of Research Publication and Reviews*, vol. 6, no. 3, pp. 18–25, 2025.

[57] P. Sundaramoorthy, R. Da, B. Puli, N. N. Jose, P. Rvs, and S. Chidambaranathan, "Ai-driven digital twin framework for personalized mental health monitoring and intervention," https://www.researchgate.net/publication/AI-driven_Digital_Twin_Framework, 2025.

[58] S. Yeasmin, S. Das, S. H. Suha, M. Prabha, N. Vanu, and A. Hosen, "Artificial intelligence in mental health: Leveraging machine learning for diagnosis, therapy, and emotional well-being," https://www.researchgate.net/publication/AI_in_Mental_Health, 2025.

[59] A. Rizwan, F. Aftab, and S. F. Abbas, "Ai-driven tools for detecting and monitoring mental health conditions through behaviour patterns," *International Journal of Psychiatry and Mental Health*, vol. 5, no. 2, pp. 8–14, 2025.

[60] M. Monfared, "Mental health counseling conversations dataset," https://www.kaggle.com/datasets/melissamonfared/mental-health-counseling-conversations-k, 2023, accessed: July 14, 2025.

[61] A. Baykal, E. F. Turetken, and M. C. Cavusoglu, "Emotion-aware dialogue systems in mental health," *IEEE Transactions on Affective Computing*, 2024, early Access.