An Intelligent and Optimized AI Framework for Mental Health Risk Identification

A

Project Report

submitted in partial

fulfillment of the

requirements for the

award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE &
ENGINEERING

By

S.no	Student Name	Roll Number	SAP ID
1	Atharv Rastogi	R2142220458	500101971
2	Rishi Raj Jain	R2142220150	500105417
3	Daksh Batra	R2142220317	500102030
4	Sunita Sapra	R2142220992	500108346

under the guidance of
Dr. Mohammad Ahsan

School of Computer Science
University of Petroleum & Energy Studies

Aug - Dec 2025

CANDIDATE'S DECLARATION

I/We hereby certify that the project work entitled **"An Intelligent and Optimized AI Framework for Mental Health Risk Identification"**in partial fulfilment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in AIML and submitted to the Department of Systemics, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **August 2025** to **December 2025** under the supervision of **Dr. Mohammad Ahsan Sir**.The matter presented in this project has not been submitted by me/ us for the award of any other degree of this or any other University.

,Daksh) Atharv, Rishi, Sunita)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 8th December 2025

Moh .Dr)amman. A(hsan

Project Guide

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Dr Mohammad Ahsan Sir** , for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank respected **Prof. Anil Kumar, Cluster head of AIML Cluster,** for his support in our projects in **Major Project 1.**

We are grateful to Dean SOCS UPES for extending requisite infrastructure for the project work .. We also thank our Course & Activity Coordinator, (Ms. Sonal Talreja) for her timely support and inputs during the project.

We herein thank all our **friends** and compatriots for their help and constructive criticism during the work.

Name	Atharv Rastogi	Rishi Raj Jain	Daksh Batra	Sunita Sapra
Enrolment No	R2142220458	R2142220150	R2142220317	R2142220992

ABSTRACT

The project **"An Intelligent and Optimized AI Framework for Mental Health Risk Identification"** proposes an AI-driven system that analyzes user interactions to detect emotional states, categorize potential mental health risk, and provide empathetic support through positive reframing. Leveraging Natural Language Processing (NLP) and transformer-based emotion classification, the framework identifies fine-grained emotions and maps them into predefined risk levels—**low, moderate, and high**—to support early awareness and timely intervention. The system is designed to be interactive and user-friendly, enabling individuals to receive real-time emotional feedback and supportive guidance, while also offering value as a supplementary tool for professionals and institutions monitoring emotional well-being. To ensure efficiency and scalability, the model incorporates optimization strategies for improved accuracy and fast inference, while explainability components such as SHAP/LIME enhance transparency in predictions and user trust. The proposed framework goes beyond conventional chatbots by combining emotion detection, risk assessment, and constructive response generation into a unified pipeline aimed at promoting self-awareness, reducing negative emotional impact, and improving preventive mental wellness support in the digital age.

TABLE OF CONTENTS

1.INTRODUCTION

Mental health has become a critical concern in today's fast-paced world, with rising stress, anxiety, and depression affecting people across age groups. Early identification and timely support are essential to prevent the escalation of emotional distress, yet traditional mental health services often face barriers such as limited accessibility, delayed responses, and lack of personalized, real-time support. These gaps create a strong need for scalable, technology-assisted tools that can complement existing care and improve early awareness.

With rapid advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP), intelligent systems can now analyze human language patterns to detect emotions and provide supportive feedback. This project, "An Intelligent and Optimized AI Framework for Mental Health Risk Identification," proposes an AI-based framework that analyzes user interactions to detect emotional states and assess potential mental health risk levels. The system is designed to work through conversational inputs and aims to function as a supportive mental wellness companion rather than a replacement for clinical diagnosis.

The core approach integrates transformer-based emotion classification—such as DistilBERT—to capture contextual meaning in user text and classify emotions using datasets like GoEmotions. Detected emotions are then mapped to structured risk categories (low, moderate, high) using a predefined risk strategy, enabling a more actionable interpretation of emotional signals. To ensure practical deployment, the framework emphasizes optimization techniques that enhance real-time inference, accuracy, and scalability.

A key differentiator of the proposed system is its focus on empathetic response generation through positive reframing. Unlike conventional chatbots that may only provide basic emotional interaction, this framework aims to support users in managing negative emotions in a constructive way. Additionally, explainability tools such as SHAP and LIME are included to improve transparency and trust by highlighting factors influencing model predictions.

Overall, this project targets a broad set of beneficiaries, including individuals experiencing emotional distress, mental health professionals seeking decision-support signals, educational institutions monitoring student well-being, and organizations promoting employee wellness. By combining emotion detection, risk assessment, explainable AI, and supportive reframing in a unified pipeline, the framework seeks to contribute toward more accessible, proactive, and awareness-oriented mental health support in the digital age.

1.1 HISTORY

The use of technology for mental health support has evolved steadily with the growth of AI and data-driven healthcare. Early digital interventions focused on basic self-help tools and rule-based systems, but the emergence of **Natural Language Processing (NLP)** expanded the possibility of understanding emotional states through text. Over time, researchers began applying **sentiment analysis** and traditional text classification methods to identify emotions such as happiness, sadness, anger, and anxiety from sources like social media, chat platforms,

and clinical text. This phase established the foundation for emotion-aware computational systems.
As the field matured, **AI-driven chatbots** gained attention for providing preliminary emotional support and coping guidance. Systems such as Woebot and Wysa highlighted how conversational agents could create accessible and immediate mental wellness interactions. However, many of these earlier systems were limited in their ability to perform **personalized risk assessment** or reliably identify more severe mental health concerns, which created a clear research gap.
The next shift came with supervised machine learning approaches designed not just to detect emotion, but to estimate psychological risk using patterns in behavior and language. This direction strengthened the case for **early intervention**, especially when paired with optimized models capable of real-time inference. Alongside this, research began emphasizing that mental health AI should not only classify negative emotion but also provide **constructive and empathetic support**, encouraging approaches like **positive reframing** and supportive response generation.
More recently, transformer-based models have become central to emotion analysis because of their strong contextual understanding. Your project aligns with this modern direction by adopting **DistilBERT** as a lightweight yet powerful emotion classifier and using datasets like **GoEmotions** to detect fine-grained emotions before mapping them into structured risk categories. This approach reflects the current trend of combining **emotion detection, risk stratification, optimization, and explainability** into a unified framework oriented toward real-time, supportive, and awareness-focused mental health assistance.

1.2 REQUIREMENT ANALYSIS

Hardware Requirements

• A computer system featuring no less than 4–8 GB of RAM

This confirms that NLP computations encompasses memory-related latency in vectorization and model loading

.• A dual-core or more powerful processor

Text preprocessing, feature extraction, and backend computations are capable of being carried out effectively with a processor that is more powerful.

• Adequate space to accommodate datasets to be maintained

Preprocessed files, model-related data, and uploaded lyrics all require storage.

Software Requirements

1. Programming Language

- **Python 3.9+** for model training, inference, and backend integration.

2. Core ML/NLP Libraries

Required for implementing the emotion detection, risk mapping, optimization, and explainability pipeline:

- **transformers**
- **datasets**
- **scikit-learn**
- **shap**
- **lime**

3. Development Environment

- **Jupyter Notebook** for experimentation and training workflows.
- **VS Code** for structured development and integration.

4. Frontend Technologies

- **Streamlit** for UI development.

5. Backend / API Layer

- **Flask or Django** for building REST APIs that handle:
  - user input requests
  - preprocessing
  - model inference
  - risk categorization
  - response delivery

6. Explainability & Visualization Support

- Integration of **SHAP** and **LIME** to generate interpretability outputs.

  Major_SRS (1)

  Major_SRS (1)
- For UI visualization of explanations:
  - **Plotly.js or D3.js**

7. Model Interface Requirements

- The ML module should use the **Transformers library** with **DistilBERT**:
  - Inputs: tokenized text and attention mask
  - Outputs: emotion logits → risk level → response text

2. MAIN OBJECTIVE

The **main objective** of this project is to **develop an intelligent and optimized AI framework** that can **detect emotions from user interactions** (as described in your reports, including text and voice), **identify potential mental health risk levels**, and **provide empathetic, supportive responses through positive reframing** in real time. The goal goes beyond basic chatbot interaction by combining **emotion detection, risk assessment, optimization for fast performance**, and **mental health awareness/early support** into a unified system.

2.1 Sub-Goals

- Build an emotion detection module using transformer-based NLP (e.g., DistilBERT) to identify fine-grained emotions from user inputs.
- Train and validate the model on a suitable dataset (GoEmotions) to ensure reliable multi-emotion classification performance.
- Design a risk assessment engine that converts detected emotions into structured mental health risk categories (low, moderate, high) through a predefined emotion-risk mapping strategy.
- Develop an empathetic response generator that provides supportive and positively reframed messages tailored to the detected emotional state and risk level.
- Optimize the model for real-time performance to enable fast emotional feedback and practical usability.
- Integrate explainable AI (XAI) using SHAP/LIME to improve transparency, interpretability, and user trust in predictions.
- Create an interactive chat-based interface that allows users to input text-based conversations and receive emotion + risk + support outputs in a simple, user-friendly manner.

- Provide a performance/evaluation dashboard to visualize accuracy, F1-score, and emotion distributions for model monitoring and analysis.
- Ensure ethical and privacy-aware design so the system remains supportive and awareness-oriented, not a substitute for clinical diagnosis.

3. SYSTEM ANALYSIS

Current technology-assisted mental health solutions—including basic emotion analyzers and AI chatbots—have improved access to preliminary emotional support. However, many existing systems still fall short in delivering **reliable risk identification**, **real-time personalized insights**, and **constructive intervention support**. This project is proposed to address these gaps by integrating emotion detection, structured risk mapping, optimization, and supportive positive reframing into a single unified framework.

3.1 Existing System

- **Limited personalization and weak risk assessment**
  Many widely referenced mental health chatbots demonstrate the usefulness of conversational support, but they often **lack personalized risk assessment** and may not accurately detect severe mental health concerns.
- **Delayed or inconsistent support in traditional systems**
  Traditional mental health services remain effective but often face **limited accessibility, delayed responses, and lack of personalized support**, making early detection difficult at scale. This creates space for AI-based supportive systems that can provide immediate awareness signals.
- **Dependence on dataset quality and generalization challenges**
  Emotion-driven AI systems may experience reduced accuracy due to **insufficient, biased, or non-diverse datasets**, affecting reliability across demographics and contexts. This is a major constraint in real-world generalization for most existing NLP-based mental health tools.
- **Difficulty in capturing nuanced emotions**
  Existing emotion detection solutions can struggle with **highly implicit or subtle emotional expressions**, especially in short or ambiguous user messages. This limits the depth of emotional understanding needed for robust risk identification.

- **Compute and scalability limitations**
  Real-time emotion and risk detection systems may require significant computational resources, which can limit large-scale deployment or accessibility on low-end systems.

- **Privacy and ethical risks**
  Handling sensitive emotional and psychological data introduces serious ethical concerns. Misinterpretation or misuse of predictions without professional oversight could lead to incorrect self-assessment or unnecessary anxiety. These concerns remain a key barrier for many existing solutions.

- **Scope limitations in many systems**
  Some frameworks rely entirely on text, which can restrict emotional accuracy when other signals (like voice tone) could help. This limitation is explicitly recognized as a weakness in current approaches.

3.2 Motivations

This project is proposed to overcome the above constraints by offering a more integrated and academically robust system design. The main motivations include:

Unified pipeline beyond basic chatbots

The framework is designed to go beyond simple conversation by combining emotion detection, risk categorization, and empathetic positive reframing, creating a more meaningful support loop for users.

Early awareness and preventive support

By enabling real-time emotional analysis and risk-level outputs, the system aims to encourage early recognition of distress patterns and promote timely support.

Optimized real-time performance

Machine learning optimization is emphasized to ensure the framework remains efficient and responsive in realistic usage conditions.

Explainable and trustworthy AI

The inclusion of explainability tools (SHAP/LIME in your SRS) strengthens transparency and user trust, helping the system justify why it inferred a particular emotional or risk outcome.

Supportive but non-clinical positioning

The framework is designed to assist awareness and emotional self-management, not to replace medical professionals—making it ethically safer and more feasible for an academic major project.

4. PROPOSED SYSTEM & MODULES

The proposed system is an intelligent and optimized AI framework designed to support mental health awareness by analyzing user conversations in a structured and meaningful way. Instead of focusing only on detecting whether a message is positive or negative, the system aims to recognize **fine-grained emotional states** expressed through the user's input. These detected emotions are then systematically interpreted and **mapped into clearly defined risk levels**, such as low, moderate, or high. This risk mapping helps transform raw emotional signals into more actionable insights, enabling the system to highlight early warning patterns of emotional distress in a way that is easier to understand and respond to.

A key strength of this framework is that it does not stop at classification. After identifying the emotional tone and risk category, the system generates **empathetic, supportive responses** that are tailored to the user's mental state. The emphasis on **positive reframing** ensures that the output is not merely informative but also constructive, helping the user view their emotions in a healthier, more balanced perspective. This approach encourages self-reflection, emotional regulation, and psychological resilience, especially for users who may not have immediate access to professional support.

Importantly, the proposed system is positioned as a **supportive mental wellness companion**, not a diagnostic or clinical tool. It is meant to assist users in understanding their emotional condition, promoting self-awareness and encouraging early intervention when needed. By maintaining this non-clinical orientation, the framework stays ethically grounded and academically feasible while still offering meaningful real-world value. In essence, the system offers a unified pipeline that combines emotion detection, risk identification, and empathetic guidance into a practical AI-based approach to promote proactive mental well-being in digital environments.

4.1 Proposed Workflow

- User provides input through a chat interface.
- Text is preprocessed and fed into a transformer model.
- DistilBERT extracts contextual embeddings and classifies input into fine-grained emotions using GoEmotions.
- Detected emotions are mapped into risk levels (Low/Moderate/High) using a predefined mapping function.
- The model is optimized for fast real-time inference with training strategies such as learning rate tuning and weight decay.
- The system generates empathetic and positively reframed supportive responses.
- Explainability outputs are produced using SHAP/LIME for transparency

4.1.1 Input Design:

The interface provides a chat-style input mechanism that allows users to enter emotional thoughts or statements for analysis. The primary input elements include:

- Input Box: Accepts user text messages or statements.

- Send Button: Submits the message to the backend and triggers the NLP pipeline for emotion and risk prediction.
- This input design is intentionally kept simple to reduce friction and encourage comfortable user expression in a mental wellness context.

4.1.2 Output Design:

After each user interaction, the system presents results in a structured, user-friendly format. The core output elements include:

- Chat Window: Displays the conversation flow, including user inputs, system-generated emotional insights, and supportive responses.
- Emotion and Risk Display: Shows detected emotion(s) with the corresponding risk level (Low, Moderate, High) to enable real-time understanding of emotional state.
- Supportive Feedback Section: Provides empathetic and positively reframed responses to encourage mental resilience and constructive emotional processing.
- Emotion Summary Panel: Presents top predicted emotions and confidence scores for clearer interpretation.
- Explanation Panel: Visualizes SHAP/LIME outputs, highlighting words or phrases that influenced predictions for transparency and trust.

4.1.3 Navigation Flow

- The navigation flow is minimal and focused to maintain a calm user experience:
- User opens the chat interface.
- User enters text into the input box and clicks Send.
- The backend processes the message and returns emotion predictions, risk level, and supportive response.
- The UI updates the Chat Window, Emotion Summary, and Risk Display.

- If enabled, the user can view the Explanation Panel to understand interpretability results.
- Users may adjust preferences using the Settings Panel (tone/language) or reset sessions.

4.1.4 Accessibility

- The UI is designed to be multi-device compatible, ensuring smooth operation on both desktop and mobile environments through responsive design principles.
- This supports broader usability for individuals with basic smartphone or computer literacy and aligns with the framework’s goal of providing accessible mental wellness support.

4.1.5 Error Handling

To maintain trust and usability in emotionally sensitive interactions, the UI should handle errors gracefully. The system should:

- Display clear messages for empty inputs, network failures, or backend service downtime.
- Provide retry options if emotion/risk outputs cannot be fetched.
- Prevent abrupt UI failures during real-time analysis.
- From an architectural standpoint, reliable communication is expected through secure REST-based request/response handling between the frontend and backend.
- The design intent is to ensure that the user experience remains stable and supportive even when technical limitations occur.

4.1.6 UI Technology Stack :

- The frontend is implemented using HTML, CSS, and JavaScript, with React or Vue recommended for a dynamic chat-based experience.
- The backend is supported by a Flask/Django REST API, and explainability visuals can be implemented using Plotly.js or D3.js.

4.2 Recommendation Engine

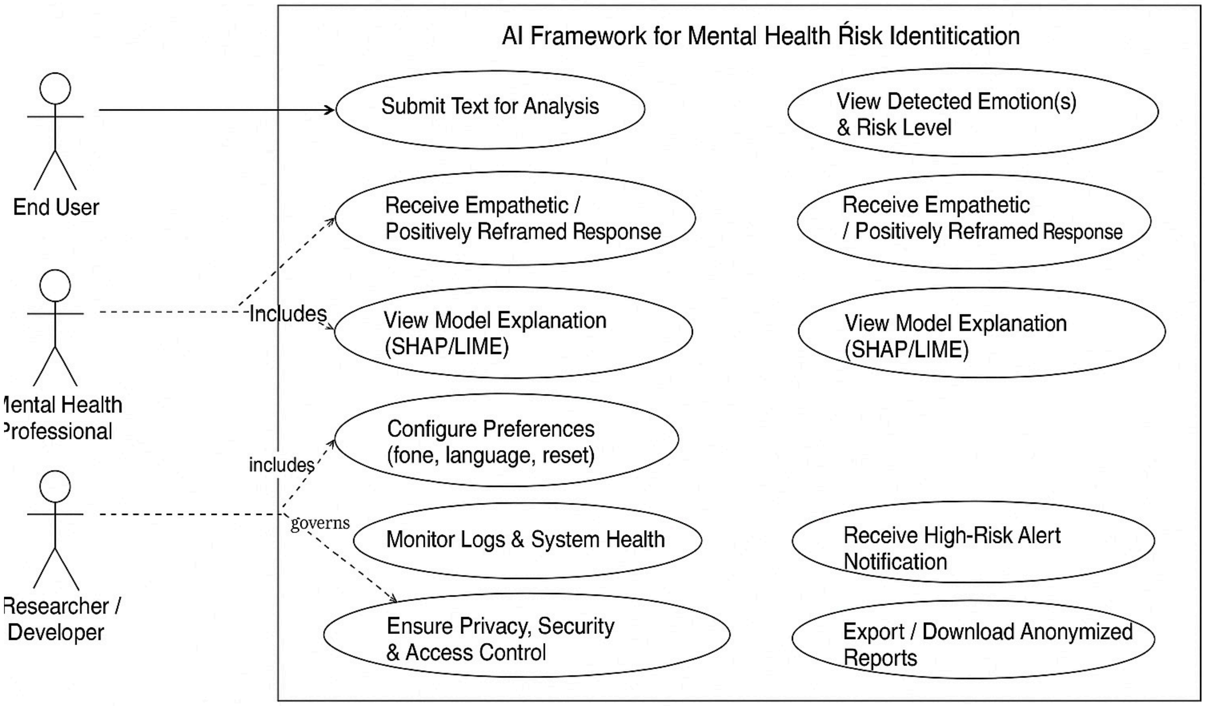
The **recommendation engine** is the part of the proposed system that provides **personalized supportive suggestions** after detecting the user's emotions and assigning a risk level. Instead of only showing an emotion label, it helps translate the prediction into **practical and empathetic guidance**, such as positive reframing messages, simple coping strategies, or calming prompts based on whether the risk is low, moderate, or high. This makes the framework more meaningful and user-focused by encouraging self-awareness and early emotional support, while still remaining a non-clinical, assistance-oriented tool.

4.3 Frontend UI

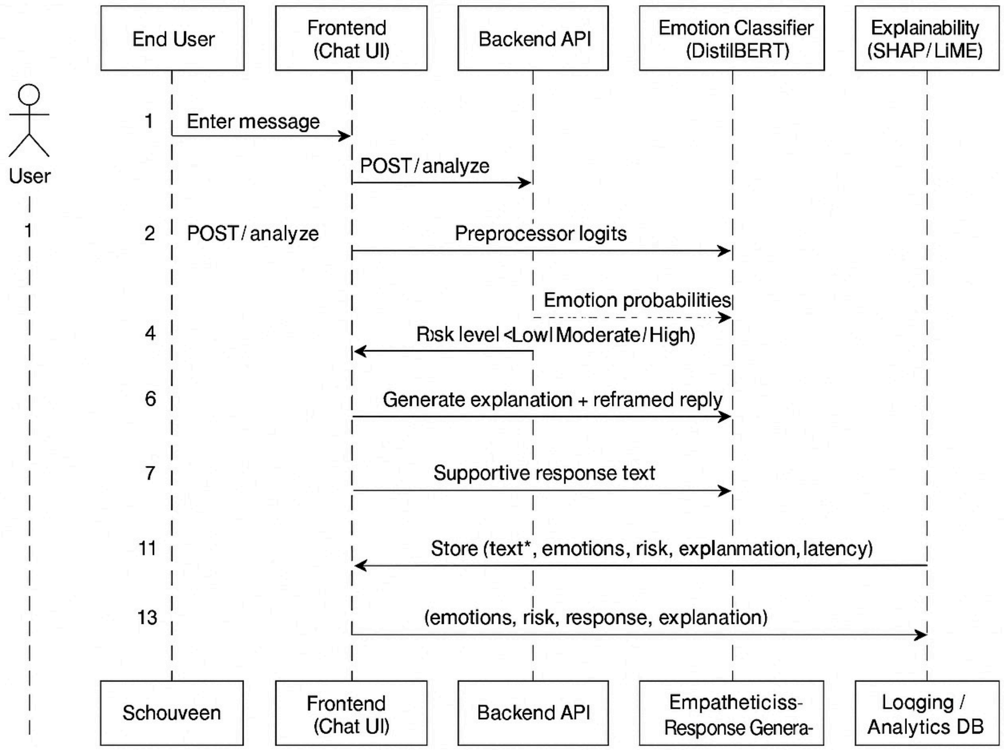
This sub-section describes the proposed frontend user interface for the AI-based Mental Health Risk Identification Framework. The UI serves as the primary communication channel between the user and the AI system and is designed to be intuitive, responsive, and empathetic, ensuring users can express themselves comfortably while receiving meaningful emotional insights in a calm and supportive manner. The interface follows a simple conversational design and supports seamless text interaction, with future extensibility for voice-based inputs.

5. DESIGN

This section consists of use cases, sequence, and Activity sketches.

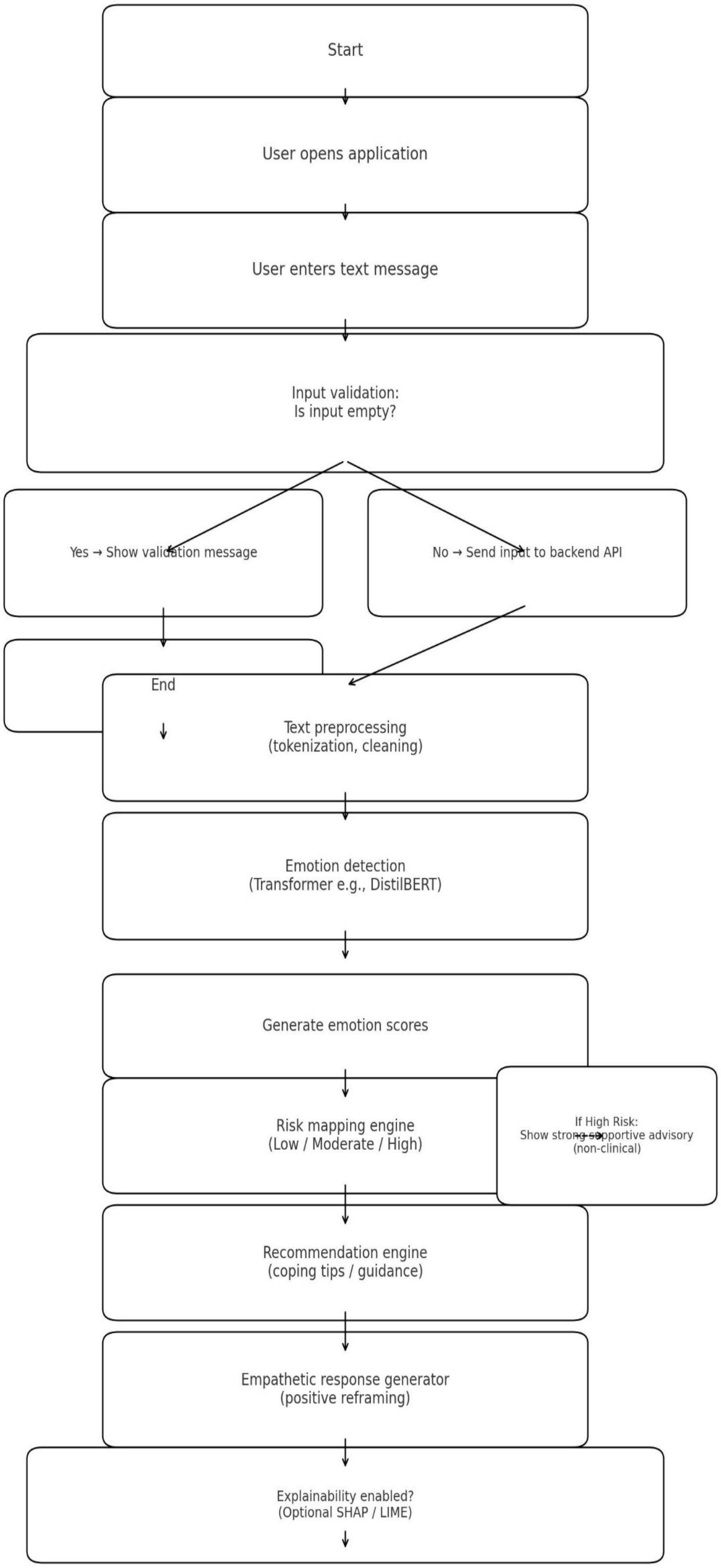

AI Framework for Mental Health Ŕisk Identitication
End User
Submit Text for Analysis
View Detected Emotion(s)
& Risk Level
Receive Empathetic /
Positively Reframed Response
Receive Empathetic
/ Positively Reframed Response
Includes
View Model Explanation
(SHAP/LIME)
View Model Explanation
(SHAP/LIME)
lental Health
Professional
Configure Preferences
(fone, language, reset)
includes
governs
Monitor Logs & System Health
Receive High-Risk Alert
Notification
Researcher /
Developer
Ensure Privacy, Security
& Access Control
Export / Download Anonymized
Reports

5.1 USE CASE


End User
Frontend (Chat UI)
Backend API
Emotion Classifier (DistilBERT)
Explainability (SHAP/LiME)
User
1
Enter message
POST/analyze
1
2
POST/analyze
Preprocessor logits
Emotion probabilities
4
Rɔsk level <Lowl Moderate/High)
6
Generate explanation + reframed reply
7
Supportive response text
11
Store (text*, emotions, risk, explanmation, latency)
13
(emotions, risk, response, explanation)
Schouveen
Frontend (Chat UI)
Backend API
Empatheticiss- Response Genera-
Loɑging / Analytics DB

1– Analyze & Support Flow

5.2 SEQUENCE DIAGRAM



5.3Activity Diagram

6. Implementation:

6.1 Overview

This chapter describes the implementation of the proposed **AI Framework for Mental Health Risk Identification**. The system is designed to analyze user conversations, detect emotional states using a transformer-based model, map emotions into structured mental health risk levels, and generate empathetic, positively reframed supportive responses. The implementation focuses on a text-centric conversational pipeline that is lightweight, scalable, and suitable for real-time user interaction.

6.2 Tools and Technologies Used

Programming Language

- Python (for ML model development, training, testing, and backend)

Libraries and Frameworks

- Hugging Face Transformers
- PyTorch
- Datasets
- Scikit-learn
- SHAP/LIME (for optional explainability)

Backend

- Flask (REST API)

Frontend

- StreamLit

6.3 Dataset Selection

To build a reliable emotion recognition system, an emotion-labeled dataset is required. A suitable dataset such as **GoEmotions** is used for training and validation. This dataset supports fine-grained multi-class emotion detection, making it appropriate for building an emotion-driven risk identification framework.

The dataset is split into training, validation, and test sets to ensure unbiased evaluation.

6.4 System Architecture Implementation

The implementation is organized into the following integrated components:

1. **Frontend UI Module**
2. **Backend/API Module**
3. **Emotion Detection Module**
4. **Risk Mapping Module**
5. **Recommendation + Response Generation Module**
6. **Explainability Module (optional)**
7. **Logging/Monitoring Module**

The frontend captures user input and communicates with the backend through a REST API. The backend calls the ML pipeline and returns results to the UI in JSON format.

6.5 Emotion Detection Module

ModelChoice
A transformer-based model such as **DistilBERT** is implemented as the core classifier due to its strong contextual understanding and lighter architecture compared to larger transformer variants.

Implementation Steps

1. Loading the pre-trained DistilBERT tokenizer and model.
2. Replacing the classification head to match the number of emotion labels.
3. Fine-tuning on the chosen dataset.

Preprocessing

- Text normalization (basic cleaning)
- Tokenization using the model tokenizer
- Padding and truncation to fixed maximum length

Output
The module produces:

- Emotion label(s)
- Confidence scores

6.6 Risk Mapping Module

This module converts predicted emotions into structured risk categories:

- **Low Risk**
- **Moderate Risk**
- **High Risk**

A predefined mapping logic is implemented to translate emotion outputs into risk labels. This ensures that emotional signals are transformed into meaningful early-awareness indicators for users.

Example Approach (Implementation Logic)

- Negative high-intensity emotions (e.g., despair-like patterns) → higher risk
- Mild negative emotions → moderate risk
- Neutral/positive patterns → low risk

This mapping is implemented as a rule-based layer for clarity, interpretability, and academic safety.

6.7 Recommendation Engine

After risk categorization, the recommendation engine selects supportive suggestions aligned with the predicted emotion and risk level. This enables the framework to go beyond detection and provide actionable guidance.

Recommendation Types

- Simple coping strategies
- Short calming prompts

- Positive reframing statements
- Motivational guidance

Implementation:
A safe and practical implementation is achieved using:

- **Rule-based selection**
- **Template-based response bank**

This ensures consistency, reduces harmful output risk, and keeps the system aligned with its non-clinical role.

6.8 Empathetic Response Generator:

This module generates emotionally sensitive, encouraging responses. The response style is designed to be supportive, respectful, and non-judgmental.

Key Design Features

- Emotion-aligned tone
- Risk-sensitive intensity
- Constructive reframing of negative thoughts

For high-risk outputs, the system can include clear supportive advisories encouraging users to seek help from trusted professionals or support systems.

6.9 Explainability Module:

To improve trust and transparency, an optional explainability layer is integrated.

Techniques Used

- SHAP
- LIME

Functionality
This module highlights influential words or phrases contributing to the model's decisions. The explanation output can be displayed inside the UI as a separate panel to enhance interpretability.

6.10 Backend/API Implementation

API Flow

1. User message is received from frontend.
2. Backend performs preprocessing.
3. Message is passed to the emotion classifier.
4. Predicted emotion scores are generated.
5. Risk mapping logic is applied.
6. Recommendation + response text is selected.
7. Optional explainability results are computed.
8. Final JSON response is sent to UI.

6.11 Frontend Implementation

The frontend is implemented as a clean chat-based interface.

Key UI Features

- Input box and send button
- Chat window for conversation history
- Emotion and risk display panel
- Recommendation and supportive response display
- Optional explainability panel

The interface is responsive to support both desktop and mobile environments.

6.12 Testing and Evaluation

Model Evaluation

- Accuracy

- Precision
- Recall
- F1-score

System Testing

- API response correctness
- UI validation for empty inputs
- Latency testing for real-time interaction
- Error handling for network and server failures

6.13 Deployment Strategy:

The system can be packaged using Docker and deployed on a cloud or local server.

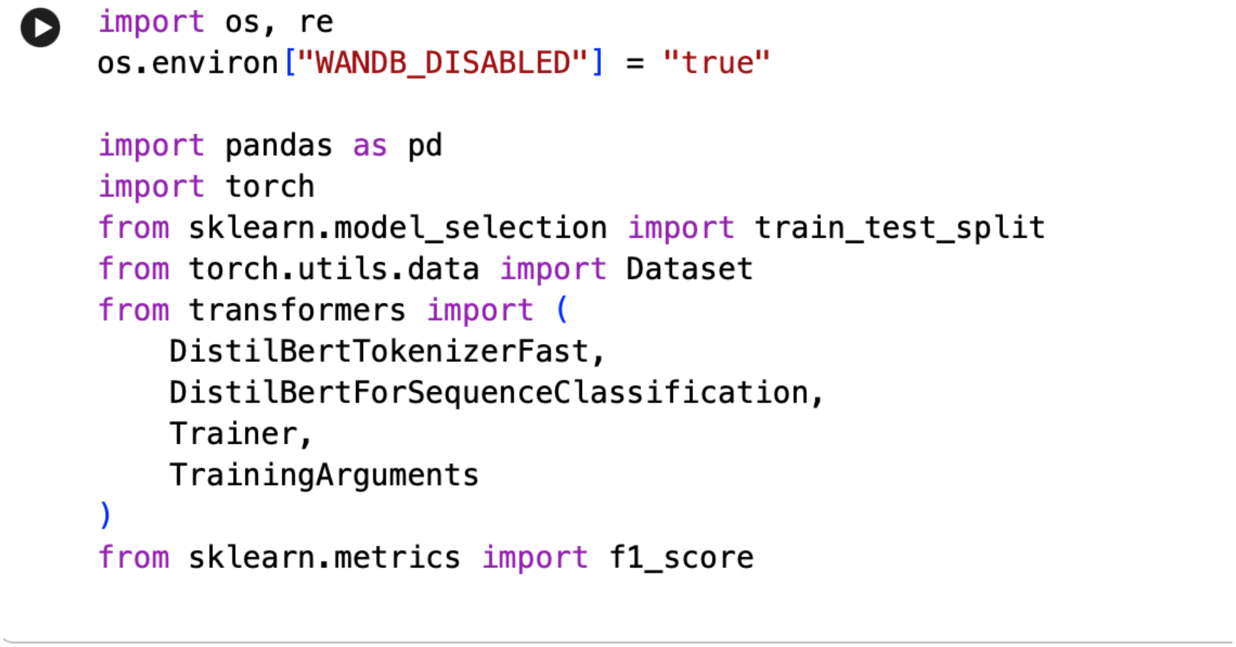
Deployment Goals

- Easy reproducibility
- Scalable real-time access
- Secure request handling

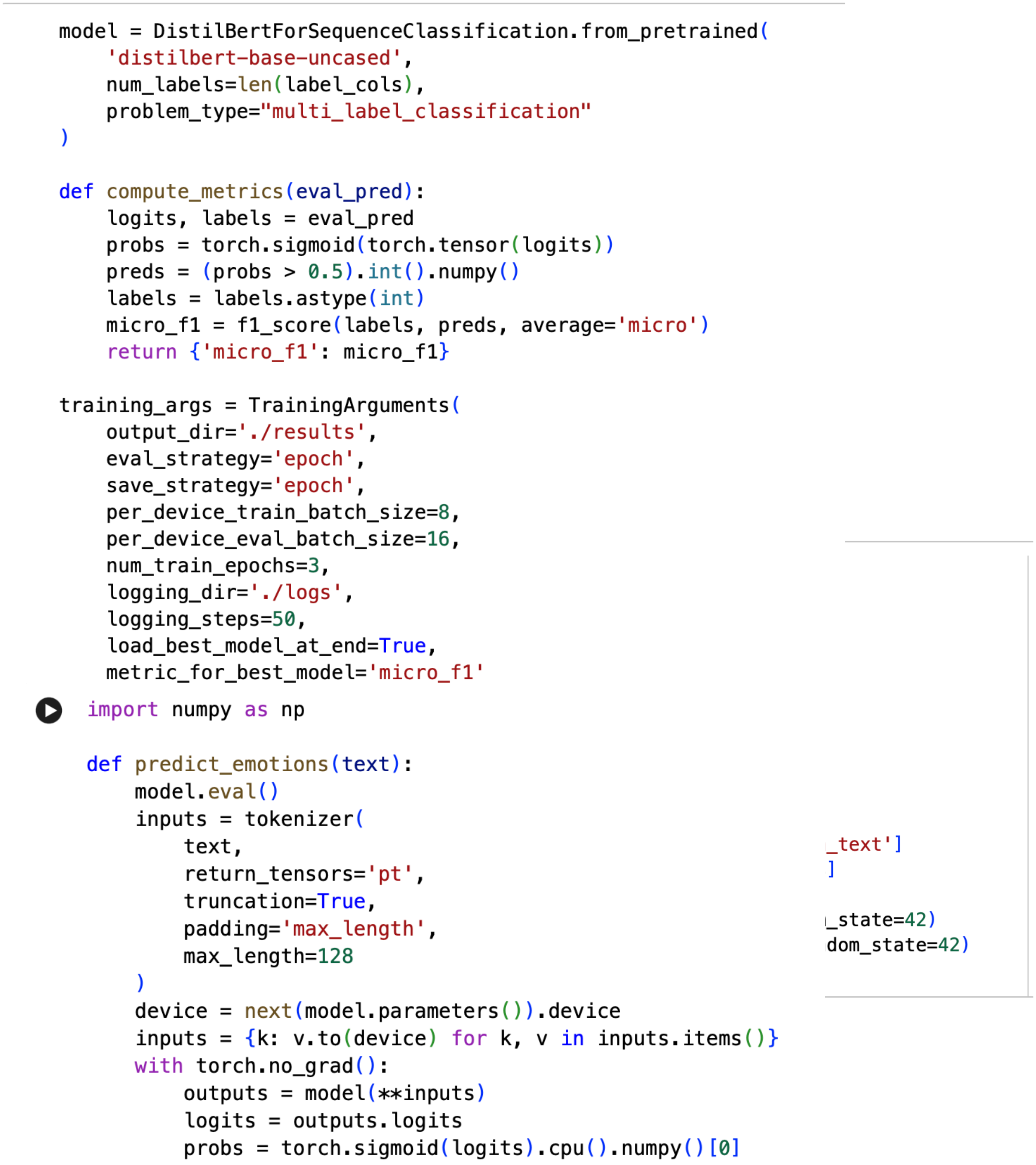
Summary 6.14

The implementation integrates a complete pipeline that combines transformer-based emotion detection, structured risk classification, and empathetic response generation. The addition of a recommendation engine ensures the framework provides practical, supportive guidance rather than only predictive outputs. With optional explainability and a user-friendly chat interface, the system achieves the objective of offering a responsible, real-time, and awareness-focused mental wellness support framework.

7. Model- Code Snippet



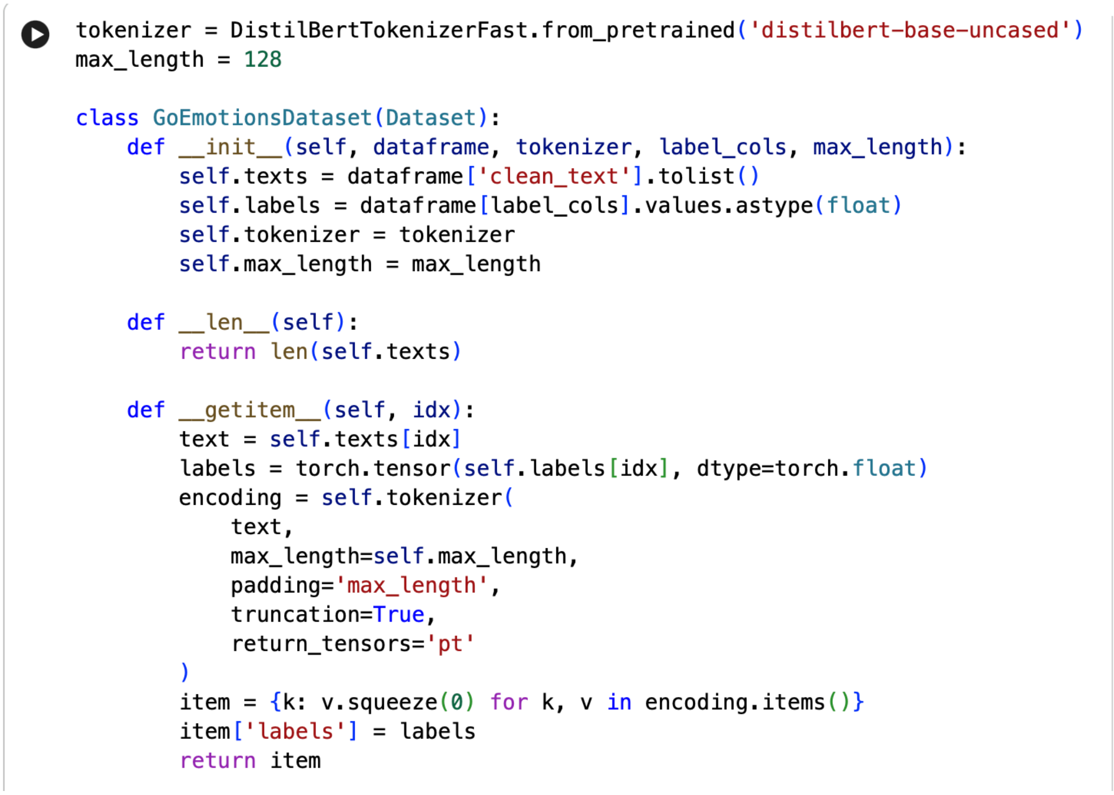
```python
import os, re
os.environ["WANDB_DISABLED"] = "true"

import pandas as pd
import torch
from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset
from transformers import (
    DistilBertTokenizerFast,
    DistilBertForSequenceClassification,
    Trainer,
    TrainingArguments
)
from sklearn.metrics import f1_score
```



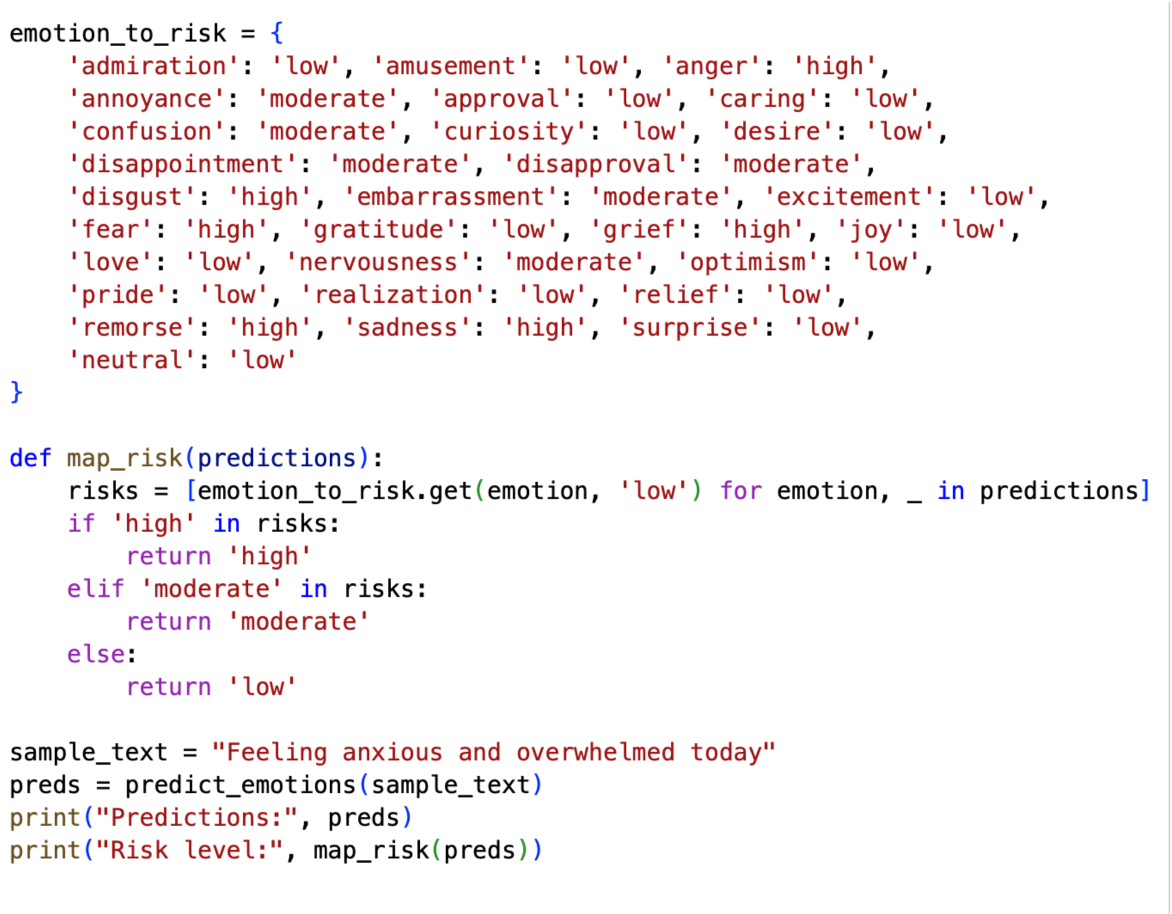
```
model = DistilBertForSequenceClassification.from_pretrained(
    'distilbert-base-uncased',
    num_labels=len(label_cols),
    problem_type="multi_label_classification"
)

def compute_metrics(eval_pred):
    logits, labels = eval_pred
    probs = torch.sigmoid(torch.tensor(logits))
    preds = (probs > 0.5).int().numpy()
    labels = labels.astype(int)
    micro_f1 = f1_score(labels, preds, average='micro')
    return {'micro_f1': micro_f1}

training_args = TrainingArguments(
    output_dir='./results',
    eval_strategy='epoch',
    save_strategy='epoch',
    per_device_train_batch_size=8,
    per_device_eval_batch_size=16,
    num_train_epochs=3,
    logging_dir='./logs',
    logging_steps=50,
    load_best_model_at_end=True,
    metric_for_best_model='micro_f1'
```

```
import numpy as np

def predict_emotions(text):
    model.eval()
    inputs = tokenizer(
        text,
        return_tensors='pt',
        truncation=True,
        padding='max_length',
        max_length=128
    )
    device = next(model.parameters()).device
    inputs = {k: v.to(device) for k, v in inputs.items()}
    with torch.no_grad():
        outputs = model(**inputs)
        logits = outputs.logits
        probs = torch.sigmoid(logits).cpu().numpy()[0]
```

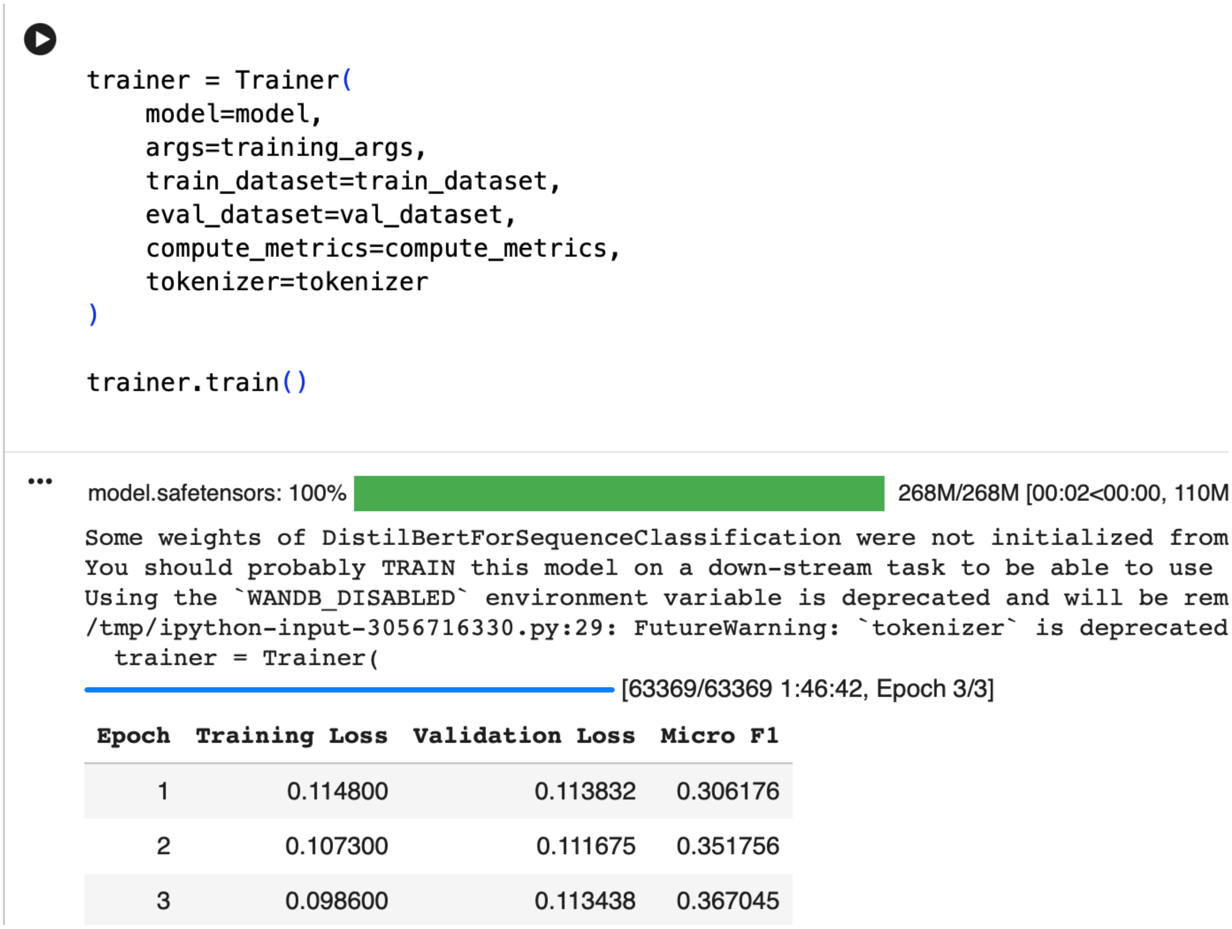
```
_text']
]

_state=42)
dom_state=42)
```



```python
tokenizer = DistilBertTokenizerFast.from_pretrained('distilbert-base-uncased')
max_length = 128

class GoEmotionsDataset(Dataset):
    def __init__(self, dataframe, tokenizer, label_cols, max_length):
        self.texts = dataframe['clean_text'].tolist()
        self.labels = dataframe[label_cols].values.astype(float)
        self.tokenizer = tokenizer
        self.max_length = max_length

    def __len__(self):
        return len(self.texts)

    def __getitem__(self, idx):
        text = self.texts[idx]
        labels = torch.tensor(self.labels[idx], dtype=torch.float)
        encoding = self.tokenizer(
            text,
            max_length=self.max_length,
            padding='max_length',
            truncation=True,
            return_tensors='pt'
        )
        item = {k: v.squeeze(0) for k, v in encoding.items()}
        item['labels'] = labels
        return item
```



```python
emotion_to_risk = {
    'admiration': 'low', 'amusement': 'low', 'anger': 'high',
    'annoyance': 'moderate', 'approval': 'low', 'caring': 'low',
    'confusion': 'moderate', 'curiosity': 'low', 'desire': 'low',
    'disappointment': 'moderate', 'disapproval': 'moderate',
    'disgust': 'high', 'embarrassment': 'moderate', 'excitement': 'low',
    'fear': 'high', 'gratitude': 'low', 'grief': 'high', 'joy': 'low',
    'love': 'low', 'nervousness': 'moderate', 'optimism': 'low',
    'pride': 'low', 'realization': 'low', 'relief': 'low',
    'remorse': 'high', 'sadness': 'high', 'surprise': 'low',
    'neutral': 'low'
}

def map_risk(predictions):
    risks = [emotion_to_risk.get(emotion, 'low') for emotion, _ in predictions]
    if 'high' in risks:
        return 'high'
    elif 'moderate' in risks:
        return 'moderate'
    else:
        return 'low'

sample_text = "Feeling anxious and overwhelmed today"
preds = predict_emotions(sample_text)
print("Predictions:", preds)
print("Risk level:", map_risk(preds))
```

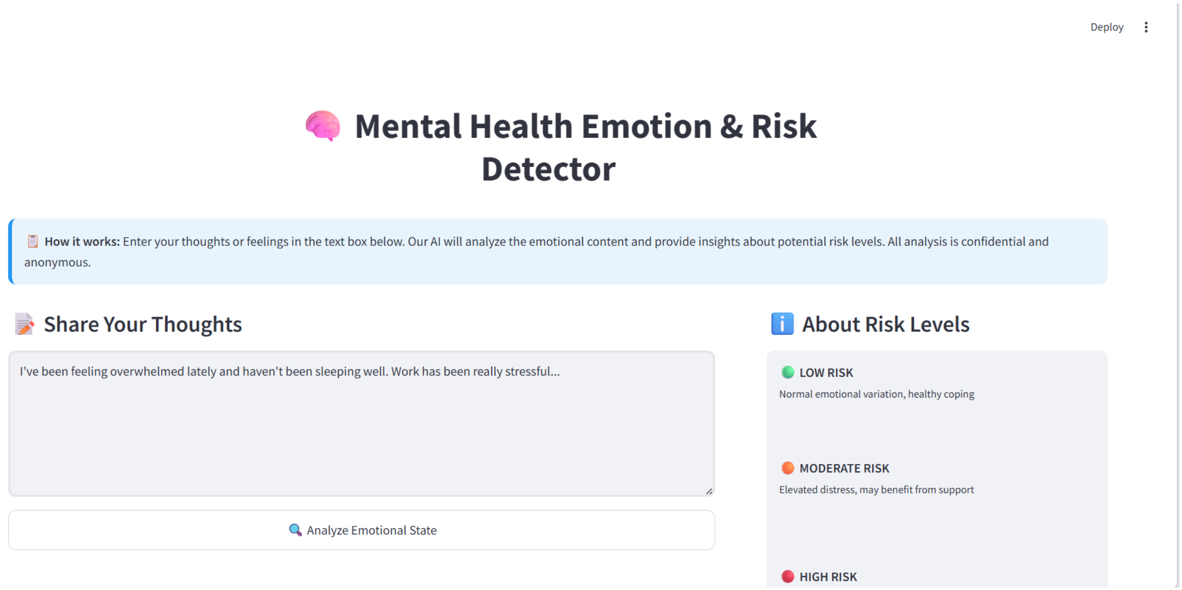


```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    compute_metrics=compute_metrics,
    tokenizer=tokenizer
)

trainer.train()
```

```
model.safetensors: 100% 268M/268M [00:02<00:00, 110M
Some weights of DistilBertForSequenceClassification were not initialized from
You should probably TRAIN this model on a down-stream task to be able to use
Using the `WANDB_DISABLED` environment variable is deprecated and will be rem
/tmp/ipython-input-3056716330.py:29: FutureWarning: `tokenizer` is deprecated
  trainer = Trainer(
[63369/63369 1:46:42, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss	Micro F1
1	0.114800	0.113832	0.306176
2	0.107300	0.111675	0.351756
3	0.098600	0.113438	0.367045

8. Output



🧠 Mental Health Emotion & Risk Detector

📋 **How it works:** Enter your thoughts or feelings in the text box below. Our AI will analyze the emotional content and provide insights about potential risk levels. All analysis is confidential and anonymous.

📝 Share Your Thoughts

I've been feeling overwhelmed lately and haven't been sleeping well. Work has been really stressful...

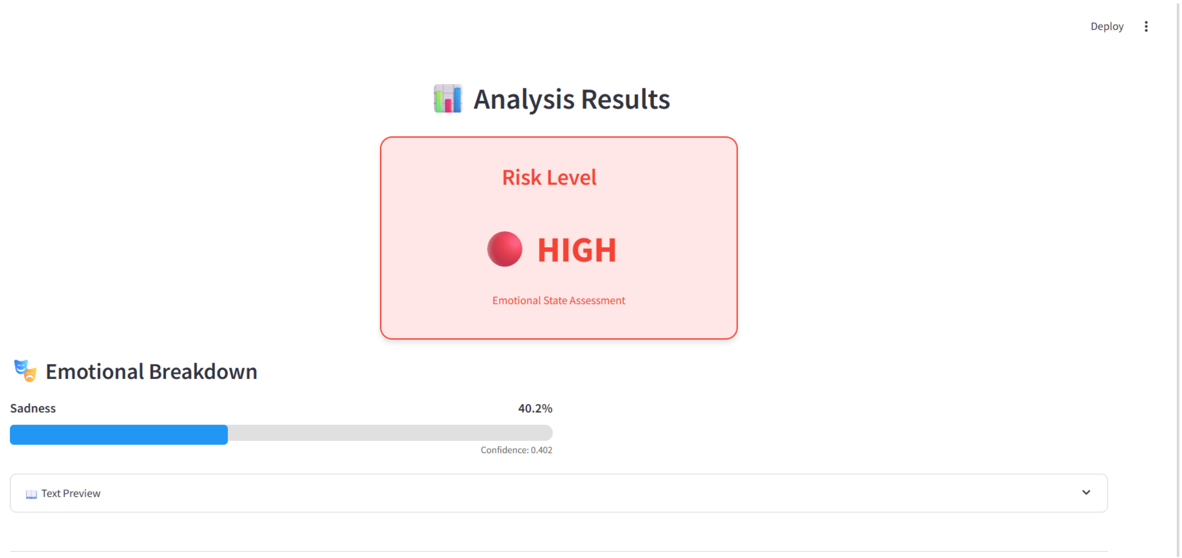
🔍 Analyze Emotional State

ℹ️ About Risk Levels

🟢 LOW RISK
Normal emotional variation, healthy coping

🟠 MODERATE RISK
Elevated distress, may benefit from support

🔴 HIGH RISK



📊 Analysis Results

Risk Level

🔴 HIGH

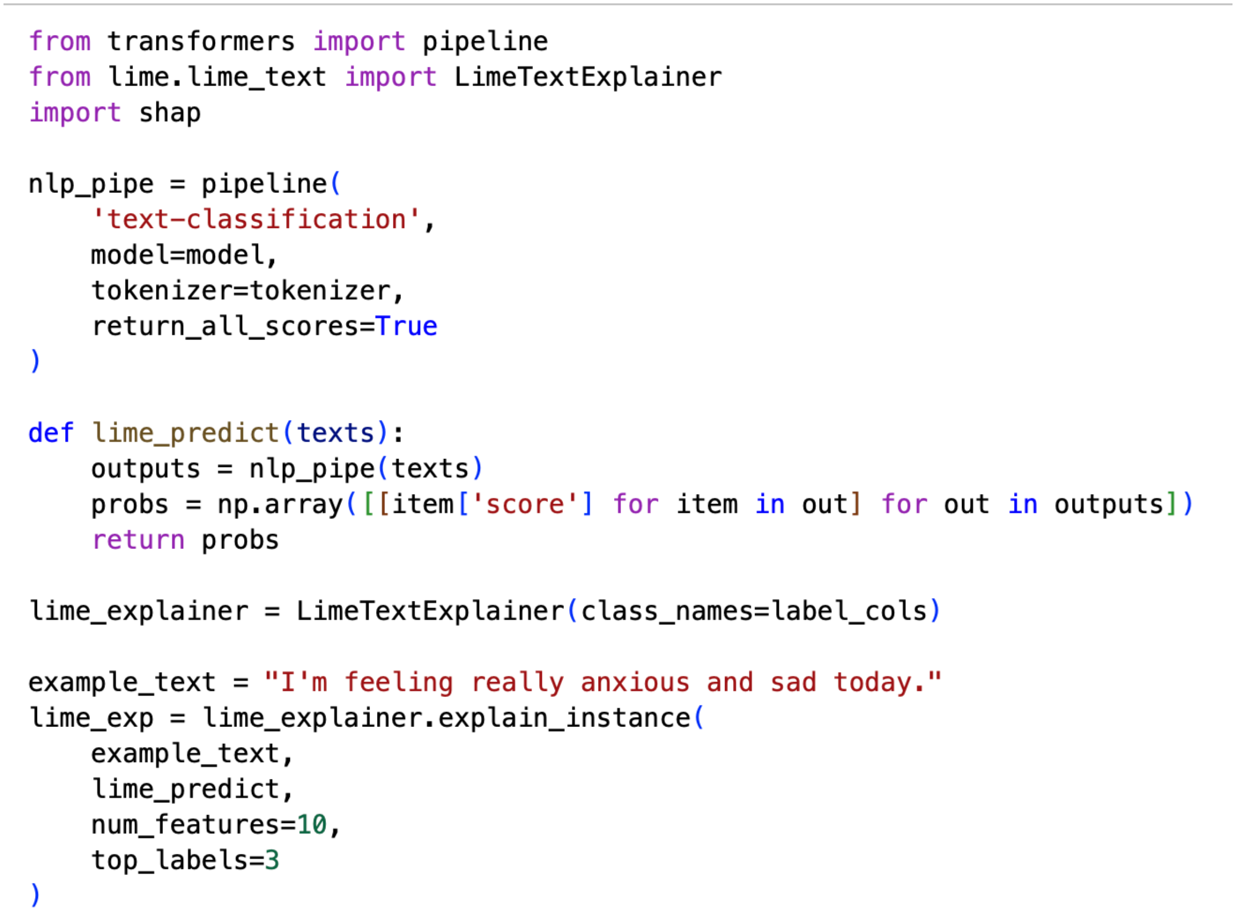
Emotional State Assessment

🎭 Emotional Breakdown

Sadness 40.2%

Confidence: 0.402

📖 Text Preview



```python
from transformers import pipeline
from lime.lime_text import LimeTextExplainer
import shap

nlp_pipe = pipeline(
    'text-classification',
    model=model,
    tokenizer=tokenizer,
    return_all_scores=True
)

def lime_predict(texts):
    outputs = nlp_pipe(texts)
    probs = np.array([[item['score'] for item in out] for out in outputs])
    return probs

lime_explainer = LimeTextExplainer(class_names=label_cols)

example_text = "I'm feeling really anxious and sad today."
lime_exp = lime_explainer.explain_instance(
    example_text,
    lime_predict,
    num_features=10,
    top_labels=3
)
```

9. Limitations

Text-dependent emotional understanding

The current implementation primarily analyzes text-based user input. This can limit accuracy when emotional cues are expressed more clearly through voice tone, facial expressions, or behavioral patterns.

Emotion ≠ clinical diagnosis

The system identifies emotions and maps them into risk levels, but this mapping is supportive and awareness-oriented, not a medically validated diagnostic measure. Emotional signals may not always reflect a person's actual clinical condition.

Dataset generalization issues

Training on public emotion datasets may not fully represent diverse cultural, linguistic, or age-specific expressions of emotion. This can affect real-world accuracy.

Subtle and ambiguous language

Short, sarcastic, or indirect messages may be misclassified because emotion is not always explicitly expressed in text.

Rule-based risk mapping constraints

If the emotion-to-risk conversion is rule-based, it may appear rigid and might not capture complex real-life emotional combinations or evolving mental states.

Response safety and depth

Supportive and positively reframed responses may still feel generic, especially for complex emotional scenarios, since the system avoids clinical advice.

Privacy and ethical sensitivity

Even with secure handling, working with mental health-related text requires careful data governance. Users may be hesitant to share personal emotions with automated tools.

Future Enhancements:

1. Multimodal emotion detection
   Extend the framework to include:

- Voice-based emotion features (tone, pitch, pauses)
- Speech-to-text integration
- (Optional) Facial expression or behavioral signals
  This can improve accuracy and make risk inference more holistic.

2. Adaptive and learning-based risk modeling
   Replace or enhance simple mapping with:

- probabilistic risk scoring
- multi-label risk inference
- dynamic thresholds based on context and confidence.

3. Improved personalization
   Add user-aware features such as:

- session-based emotion trends optional mood history
- personalization of tone and recommendation style.

4. Stronger recommendation library
   Expand the recommendation engine to include:

- structured coping plans
- guided self-help routines
- daily emotional check-in prompts
- stress management and mindfulness micro-exercises.

5. Safer and richer response generation
   Introduce a hybrid approach:

- curated templates for safety
- optional controlled AI generation with strict guardrails
  to make responses more natural but still responsible.

6. Bias and fairness evaluation
   Conduct broader testing across:

- languages
- demographic groups
- regional and cultural emotion patterns
  to ensure consistent performance.

7. More detailed evaluation and benchmarking
   Add:

- baseline comparisons (traditional ML vs transformers)
- ablation studies
- error analysis by emotion class
  to strengthen academic reliability.

8. Deployment and scalability improvements
   Enhance real-world readiness by adding:

- containerized deployment
- load testing
- caching and model quantization
  for faster inference under higher traffic.

9. Crisis-handling and escalation guidance
   For high-risk outputs, add a carefully worded safety layer that:

- encourages reaching out to trusted people
- suggests professional support
- provides region-appropriate emergency guidance
  without claiming medical authority.

10. CONCLUSION

This project presented an **intelligent and optimized AI framework for Mental Health Risk Identification** that focuses on understanding user conversations, detecting emotional states, and transforming these signals into structured risk categories. By combining **transformer-based emotion classification**, a **risk mapping strategy**, and a **recommendation and positive reframing mechanism**, the system moves beyond simple emotion detection and offers supportive, awareness-oriented guidance in real time. The inclusion of an interactive frontend and a backend API-based architecture further strengthens the feasibility of deploying the framework as a practical mental wellness assistance tool.

A key contribution of this work is its attempt to bring together **emotion analysis, early risk indication, empathetic response generation, and optional explainability** in a unified pipeline. This helps improve user trust and makes the outputs more interpretable and meaningful. At the same time, the system is positioned ethically as a **supportive companion** rather than a clinical diagnostic solution, ensuring responsible use in sensitive mental health contexts.

Overall, the proposed framework demonstrates how AI can be applied to promote **self-awareness, early emotional intervention, and constructive coping support**. With further enhancements such as multimodal inputs, improved personalization, and advanced risk modeling, the system has strong potential to evolve into a more comprehensive and high-impact mental health support platform.

11. REFERENCES

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

- Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth and uHealth, 11*, e44838.

- Oghenekaro, L., & Okoro, C. (2024). Artificial Intelligence-Based Chatbot for Student Mental Health Support. *Open Access Library Journal, 11*, Article 1511

- Lee, S., Park, J., & Kim, Y. (2025). Development and Evaluation of a Mental Health Chatbot Using DSM-5 Criteria and a Korean Corpus. *JMIR Medical Informatics, 13*, e63538.

- Booth, F., Potts, C., Bond, R., Mulvenna, M., Kostenius, C., Dhanapala, I., Vakaloudis, A., Cahill, B., Kuosmanen, L., & Ennis, E. (2023). A Mental Health and Well-Being Chatbot: User Event Log Analysis. *JMIR mHealth and uHealth, 11*, e43052.

- Booth, F., Potts, C., Bond, R., Mulvenna, M., Kostenius, C., Dhanapala, I., Vakaloudis, A., Cahill, B., Kuosmanen, L., & Ennis, E. (2023). A mental health and well-being chatbot: User event log analysis. *JMIR mHealth and uHealth, 11*, e43052.

- Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine, 5*, Article 46.

- Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. *Frontiers in Digital Health, 6*, Article 1280235.

- Balcombe, L. (2023). AI Chatbots in Digital Mental Health. *Informatics, 10*(4), 82 \

- Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI, 2*(4), Article AIoa2400802