# AUDIO-DRIVEN EMOTION RECOGNITION AND CONTEXTUAL SUMMARIZATION FOR MENTAL HEALTH CONVERSATIONS

Team Members: Ravi Raj(22bds051), Parishri Shah(22bds043), Pakhi Singhal(22bds042) and Preethi Varshala S (22bds045)

# Introduction

Global Impact of Mental Health: Millions of individuals are affected by mental health disorders worldwide, necessitating accessible and effective support systems.

**Two key AI-driven approaches:**

- **Audio Driven Emotion Recognition (ADER)**
- **Contextual Summarization and Emotion Prediction (CSEP)**

# Dataset Overview

**Dataset Size:**

- 5,247 rows are used for emotion detection from text using BERT.
- 1,163 rows are used for emotion detection from audio using MFCC features.

**Key Features:**

- Utterance: Text of each conversation line.
- Emotion Labels:
  - -1: Negative emotion.
  - 0: Neutral emotion.
  - 1: Positive emotion.
- Dialog Acts: 55 categories, capturing conversation functions (Information-Request, General Chit-Chat)

# Audio Driven Emotion Recognition (ADER)

Step 1: Converts the text-based utterances into speech (audio) files (e.g., .mp3) using Google Text-to-Speech (GTTS).

Step 2: MFCC Feature Extraction

Step 3: Machine Learning Model Training:

- Training Process:
  - MFCC features are fed into two machine learning models - XGBoost and Random Forest.
  - The model is trained to classify emotions as negative (-1), neutral (0), or positive (1).
- Dataset Split: The data is split into training and testing sets to ensure proper evaluation of model performance.

Step 4: Evaluation

- XGBoost - 77% accuracy
- Random Forest - 75% accuracy

# ADER

## Prediction Workflow

1. Input Processing:

- If the input is text, it is first converted to audio using GTTS.
- If the input is already in audio format, this step is bypassed.

2. Feature Extraction:

- MFCC features are extracted from the audio file.

3. Emotion Prediction:

- The extracted features are fed into the trained model, which predicts the emotional state (negative, neutral, or positive).

# Contextual Summarization and Emotion Prediction (CSEP)

Step 1: Preprocessing - Tokenization of Utterances, label Encoding of Dialogue Acts and Emotion columns and using PyTorch, tokenized text is embedded into numerical representations that capture contextual meaning. Divide the dataset into training, validation, and test sets.

Step 2: Model Training - Input features are embedded representations of the text. Output labels are encoded emotions. Use the custom BERT model for training. K-fold cross-validation is applied during training to ensure robust model evaluation.

Step 3: Model Evaluation - 94% Accuracy

# CSEP

## Prediction Workflow

1. Input Text Check:

- If the sentence has more than 30 words, it proceeds to text summarization. Text Summarization is done using BART.
- If the sentence has 30 words or fewer, it skips summarization.

2. Emotion Detection:

- Whether the original or summarized text, it is processed using a BERT model for emotion prediction.
- The model predicts one of three emotional tones: negative, neutral, or positive.

# Key Achievements and Conclusions

- AI-Driven Emotion Detection:
  - Robust framework for detecting emotions using speech and text data.
  - The ADER workflow effectively classifies emotional states based on MFCC features.
- Contextual Summarization:
  - The CSEP system enables meaningful summarization, preserving emotional context and reducing conversational complexity.After summarization, we used BERT to detect emotions from the text in the utterances, ensuring accurate emotion classification.
- Overall Impact:
  - This AI-driven system enhances patient-therapist interactions by providing real-time emotion detection from both text and audio and summarization, helping to improve mental health support systems.

# THANK YOU

-