Voice2sentiment: An End – to – End System for Speech Emotion Recognition and Textual Sentiment Analysis

UNDER THE GUIDANCE OF

Prof. Dr. Dhrubasish Sarkar & Sonali Das

Department of B.Tech in Computer Science Engineering

(Specialization in Artificial Intelligence and Machine Learning)

Supreme Knowledge Foundation Group of Institutions, Hooghly, West Bengal

Under

Maulana Abul Kalam Azad University of Technology Kolkata, West Bengal, India.

SUBJECT : PROJECT -I

SUBJECT CODE : PROJ-AIML781

TEAM MEMBER	UNIVERSITY ROLL NUMBER
Uma Saha	25330822021
Jamima Khatun	25330822009
Debolina Samanta	25330822008
Aditya Choudhary	25330822026

ABSTRACT

- **Motivation**

➢ Most systems analyze either voice tone or text meaning — rarely both. Our project combines speech emotion recognition with sentiment analysis to better understand feelings and context, making human–machine interaction more natural.

- **Background/Context**

➢ Many voice assistants and chatbots understand our words but not the feelings behind them. Most do either speech-to-text or emotion detection, not both — and often only in one language. Voice2Sentiment does both, in Bengali and English, so machines can respond more naturally.

- **Summary of Problem Statement**

➢ Our main motive is to build an end-to-end system that can analyze both voice tone and text transcription with sentiment analysis with some better accuracy.

INTRODUCTION

- **Problem Statement**

Most systems check voice or text separately and don't combine tone and spoken words in real time. They also usually work in only one language, not multiple like Bengali and English.

- **Gap Analysis**

Lack of unified real-time systems combining speech-to-text, emotion recognition, and sentiment analysis. Existing models often fail in noisy environments and cross-modal emotion interpretation.

- **Purpose of the Project**

To develop an end-to-end intelligent system that captures human voice, converts it to text, and analyzes both for emotional and sentiment context using Artificial Intelligence and Machine Learning techniques.

- **Scope of the Project**

Voice capture & noise reduction, speech-to-text via ASR, audio feature extraction(MFCC,pitch,chroma),NLP –based sentiment analysis, emotion classification using ML(SVM,ANN,HMM)and secure storage and real-time response.

- **Significance**

Enables multimodal emotion detection for use in mental health, virtual assistants, and AI-driven communication – offering a scalable, intelligent, and human-aware interaction systems.

OBJECTIVES

SUPREME KNOWLEDGE FOUNDATION

- Listen to the voice and read the words at the same time.
- Understand the actual feeling behind what is said.
- Detect if a positive sentence is said in a negative or sarcastic way.
- Find emotions like happiness, anger, sadness, sarcasm, or mockery.
- Compare the meaning of words with the tone of voice.
- Give the true sentiment, not just what the words mean.
- Work for multiple languages and accents.
- Help in real-life cases like customer service, social media, or interviews.

RESEARCH GAP

- Most systems check only the words or only the tone, not both together.
- They miss sarcasm or hidden feelings when tone and words are different.
- Few systems can compare meaning of words with the way they are spoken.
- Many do not work well with different languages, accents, or background noise.
- Real-time, accurate sentiment with both voice and text is still rare.

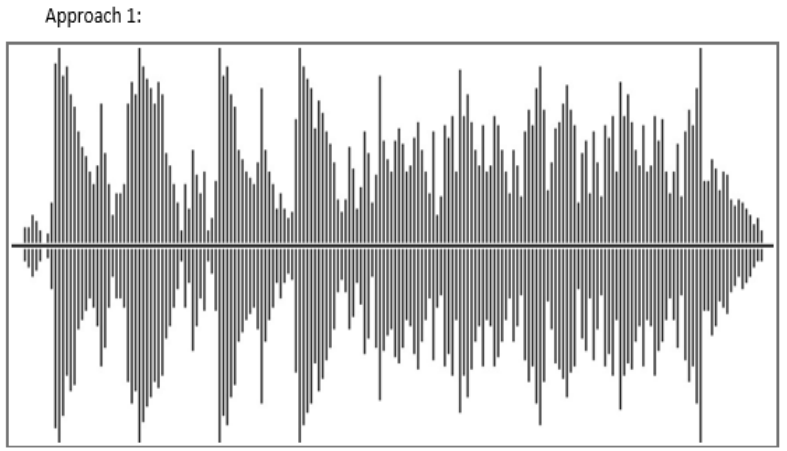
Proposed Solution

SUPREME KNOWLEDGE FOUNDATION

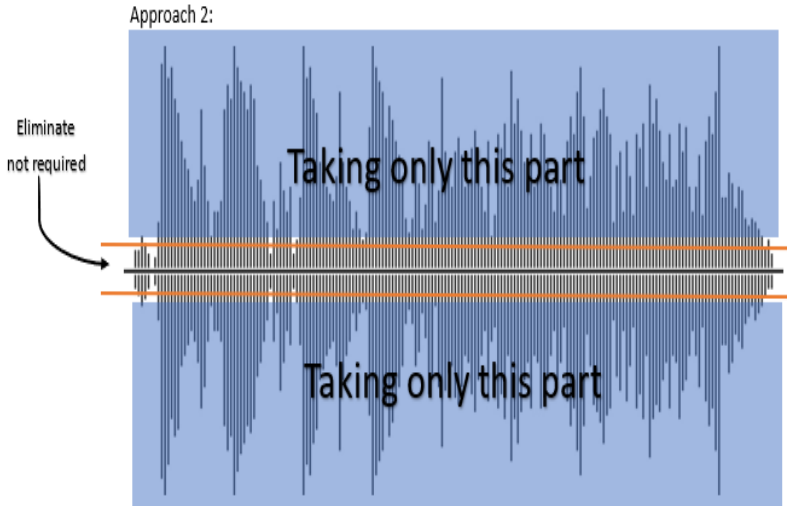
❖ **Detailed explanation of the proposed solution**

Solution 1 : Build Our own Deep Learning model for analyzing Emotion there are following two way to get data from voice data :

- **Approach 1:** Use raw extracted audio feature arrays as input for emotion classification.
- **Approach 2:** Define decision boundaries or range zones to minimize overfitting and complexity, which typically arise due to normalization and noise in real data. This addresses the **overfitting** and **high time complexity** problems often found in standard models.



Solution 2: Using a pre-trained LLM like Qwen2.5 Omni- 3B requires only the data. You can use it directly without fine-tuning, since it's already trained on a large dataset. Fine-tuning is only needed for custom tasks, where we adjust the model's weights to specialize in a specific kind of work. Both approaches are valid.



Solution 3: Combine both Deep Learning and LLM Models

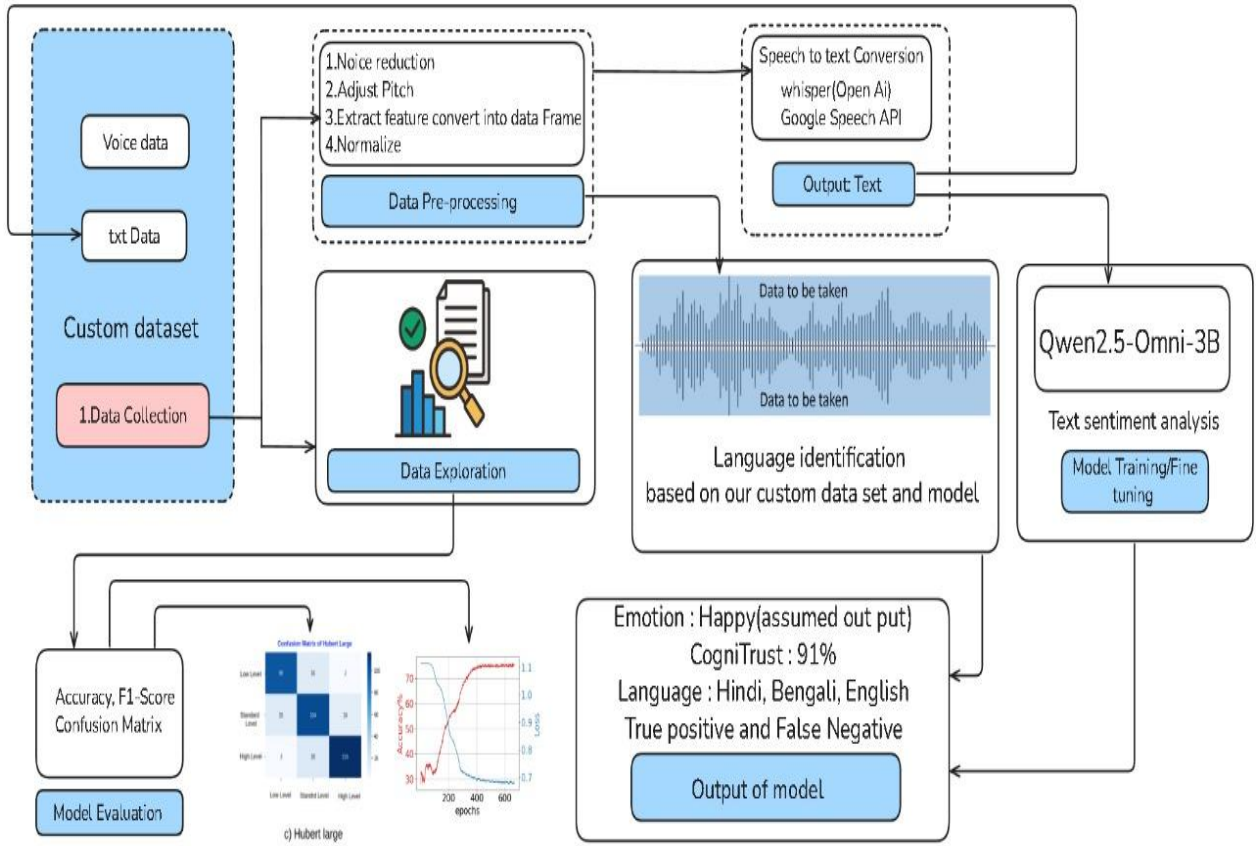
TECHNICAL APPROACH

SUPREME KNOWLEDGE FOUNDATION

- **Technologies to be used :**

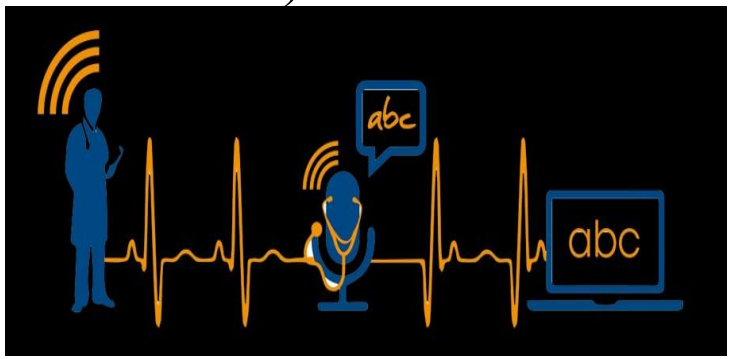
•***Languages***: Python (core), JavaScript (backend/frontend),Django or Flask(backend),HTML&CSS(frontend)
•***Frameworks/Libraries***:
•PyTorch / TensorFlow (deep learning)
•Hugging Face
•Whisper Speech Recognition (speech-to-text)
•Lib ROSA / PyDub (audio processing)
•VADER / TextBlob (sentiment analysis)
•***Tools***: Streamlit, Flask, Google Colab, GitHub
•***Hardware***: *8*GB+ RAM system, GPU recommended, or cloud (AWS/GCP),render.

- **Methodology and process for implementation:**



FEASIBILITY AND VIABILITY

- **Analysis of the feasibility of the idea**
  - ❖ Technically possible using existing tools(Whisper, Qwen2.5 Omni- 3B etc.).
  - ❖ Operational on local or cloud platforms.
  - ❖ Cost-effective with open-source models.
  - ❖ High relevance in real-world applications.



- **Potential challenges and risks**
  - ❖ Emotion ambiguity in speech .
  - ❖ Background noise affecting accuracy.
  - ❖ Limited or biased datasets.
  - ❖ Real-time latency issues.
  - ❖ Data privacy concerns.

- **Strategies for overcoming these challenges**
  - ❖ Use ensemble /multi-modal models, apply noise reduction techniques and train on diverse, augmented datasets.
  - ❖ Optimize models for fast inference.
  - ❖ Ensures privacy via encryptions and GDPR compliance.

IMPACT AND BENEFITS

- **Potential impact on the target beneficiary.**
  - Enhances global customer care through emotionally aware responses.
  - Empowers early mental health support, reducing long-term medical burden.
  - Makes virtual assistants more human-like and emotionally intelligent.
  - Create safer, more respectful online environments.
  - Enables businesses to make smarter, data-driven marketing and product decisions.
- **Benefits of the solution (social, economic, environmental, etc.)**
  - Social: Improves mental health detection, moderates harmful content, and supports more empathetic technology.
  - Economic: Accelerates market research and enhances customer services, boosting brand value and operational efficiency.
  - Environmental: Promotes ethical AI use through natural voice interface and leverages sustainable, open-source tools.
  - **Business Purpose :**
  - Real-Time Insights: Understand emotions and sentiments in speech and text instantly.
  - Enhanced Decision-Making: Help organizations act on customer, employee, or audience mood intelligently.
  - Customer & Employee Experience: Detect satisfaction, frustration, or distress for proactive engagement.
  - Mental Health & Safety: Monitor emotional well-being in healthcare, counseling, and security contexts.
  - Brand & Reputation Management: Track public sentiment, social media trends, and potential crises.

REFERENCES

- Sehgal RR, Agarwal S, Raj G. Interactive voice response using sentiment analysis in automatic speech recognition systems. In2018 International Conference on Advances in Computing and Communication Engineering (ICACCE) 2018 Jun 22 (pp. 213-218). IEEE.
- Rao A, Ahuja A, Kansara S, Patel V. Sentiment analysis on user-generated video, audio and text. In2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) 2021 Feb 19 (pp. 24-28). IEEE.
- Anand S, Patra SR. Voice and text based sentiment analysis using natural language processing. InCognitive Informatics and Soft Computing: Proceeding of CISC 2021 2022 May 31 (pp. 517-529). Singapore: Springer Nature Singapore.
- https://www.sciencedirect.com/topics/computer-science/speech-emotion-recognition
- https://huggingface.co/firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3
- https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition
- https://medium.com/heuristics/audio-signal-feature-extraction-and-clustering-935319d2225