ASPECT BASED SPEECH EMOTION RECOGNITION

Deep Learning Course Project Final Report
2025

Student 1: Abdul Ahad	**CMS: 023-22-0206**
Student 2: Suman	**CMS: 023-22-0138**
Student 3: Qurban Ali	**CMS: 023-21-0166**

SUKKUR IBA UNIVERSITY
Computer Science Department
(2025)

Table of Contents

Abstract

This project introduces a novel Aspect-Based Speech Emotion Recognition (AB-SER) framework tailored for the food domain. Unlike traditional systems that handle speech emotion or aspect sentiment separately, the proposed model integrates both acoustic emotion features and aspect-level textual analysis to generate deeper, context-aware insights from spoken feedback. The system utilizes accurate models like Wav2Vec2 for emotion detection and Aspect-Based Sentiment Analysis (ABSA) for identifying specific feedback targets, offering potential for real-world deployment in the food industry.

1.0 Introduction

Speech Emotion Recognition (SER) is a growing field that uses deep learning to identify human emotions from speech signals, typically classifying states such as happiness, anger, or sadness using acoustic features like pitch and tone [1]. However, traditional SER does not capture context especially in domain-specific settings such as food-related customer interactions. To overcome this limitation, recent research combines SER with Aspect-Based Sentiment Analysis (ABSA), which identifies both the emotion and the specific aspect being discussed (e.g., taste, pricing, service) [2].

In the food industry, large amounts of spoken feedback are collected via customer support calls, voice assistants, and review recordings, yet most analysis remains text-based. Aspect-level emotion detection can help restaurants and food services automate complaint handling, monitor satisfaction, and improve service quality by identifying which aspect triggers which emotion.

This project proposes an Aspect-Based Speech Emotion Recognition (AB-SER) model that integrates speech emotion classification with aspect extraction from transcribed text, enabling fine-grained analysis of emotional opinions in the food domain.

2.0 Problem Identification

2.1 Background

Existing speech emotion recognition systems classify only global emotions without considering what the emotion refers to [3]. Similarly, aspect-based sentiment systems work only on text and ignore acoustic emotional signals. There is currently no integrated system that performs aspect-level emotion detection from speech in the food domain.

The problem addressed in this project is therefore: How can we automatically detect emotions from spoken food reviews and map them to specific aspects such as taste, delivery, service, or hygiene, using a combination of speech and text-based deep learning models?

This problem is relevant to food businesses, customer-care platforms, and food delivery apps that receive large-scale voice feedback but lack automated analysis tools. Solving it will enable improved customer experience, automated complaint triaging, and data-driven decision-making in the food industry.

2.2 Research Questions

1. How can speech emotion recognition and aspect-based sentiment analysis be combined to generate aspect-level emotional insights from spoken food reviews?
2. Can multimodal fusion of acoustic emotion features (from audio) and semantic aspect-sentiment features (from text) improve classification performance compared to unimodal systems (audio-only or text-only)?

2.3 Hypothesis

It is hypothesized that:

1. **Multimodal Superiority Hypothesis:** A multimodal deep learning model that jointly uses acoustic speech features and textual aspect-sentiment features will achieve significantly higher performance than unimodal models (audio-only or text-only) for aspect–emotion classification.
2. **Emotion–Aspect Dependency Hypothesis:** The emotional intensity detected from speech (e.g., angry tone, happy tone) is correlated with the sentiment polarity of the corresponding aspect in the transcript (e.g., negative tone → negative sentiment on “service” aspect).
3. **Domain Adaptation Hypothesis**: Fine-tuning pre-trained models (Wav2Vec2 and BERT) on an additional domain-specific spoken food review dataset, after training on benchmark datasets (RAVDESS and SemEval), will improve performance compared to using the benchmark datasets alone.

2.4 Contributions of the Project

1. **Multimodal AB-SER Framework:** Introduces the first integrated model that combines speech emotion recognition with aspect-based sentiment analysis to extract aspect–emotion pairs from spoken food reviews.
2. **Benchmark vs Fine-Tuned Performance Study:** Provides a comparative evaluation showing how multimodal fusion and domain fine-tuning improve accuracy over traditional unimodal and benchmark-only models.

3.0 Objectives

The primary objectives of this project are as follows:

1. Build a multimodal model combining speech + text
2. Detect emotion from audio
3. Extract aspects from transcript
4. Classify sentiment polarity for each aspect
5. Fuse both to produce aspect → emotion mapping

4.0 Dataset Discussion

Since no publicly available dataset exists specifically for emotion-annotated food speech, this project adopts a separate dataset strategy using two freely available benchmark datasets:

COMPONENT	DATASET	DOMAIN	USAGE
Speech Emotion	RAVDESS	General speech	Train & evaluate SER model
Text ABSA	SemEval-2014 Task 4 (Restaurant reviews)	Food	Train aspect extraction + sentiment polarity model

4.1 Dataset Preparation

The system processes raw audio inputs before they are fed into the model.

The preprocessing steps implemented include:

- **Resampling:** All audio files are resampled to 16,000 Hz (sr=16000) to match the input requirements of the Wav2Vec2 model.
- **Cleaning:** Corrupted files are identified and skipped. In cases of partial corruption, silence handling (zero-padding) is applied.
- **Padding/Truncation:** Audio inputs are padded or truncated to a fixed maximum length (e.g., 5 seconds) to ensure uniform batch processing.
- **Label Mapping:** Categorical emotion labels (e.g., “neutral”, “calm”, “happy”) are mapped to integer IDs for classification.

4.2 Previous Benchmark Results

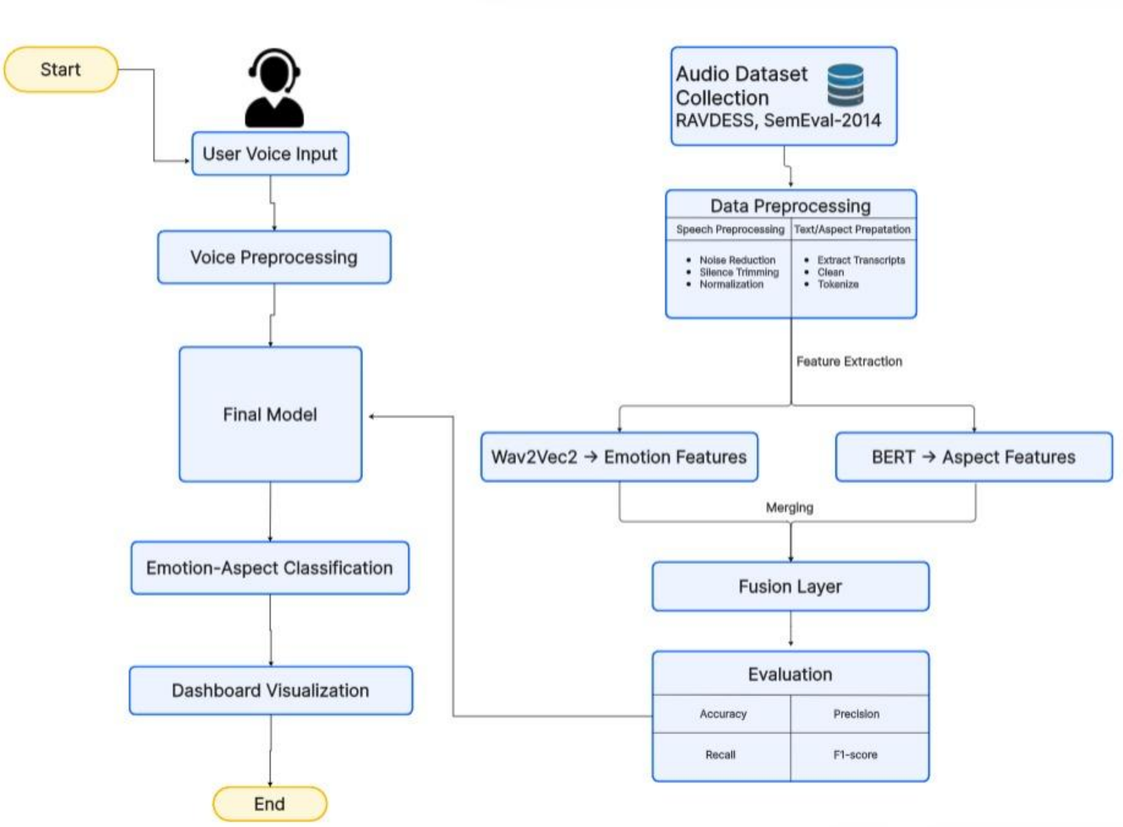
RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a validated dataset widely used in the research community. Previous benchmarks on RAVDESS using CNN and RNN architectures typically achieve accuracies in the range of 60-80% depending on the complexity of the feature extraction (MFCCs vs. raw waveform). State-of-the-art results using transformer-based models like Wav2Vec2 have pushed this boundary higher, often exceeding 80% accuracy.

4.3 Rationale for Dataset Selection

RAVDESS provides high-quality, actor-recorded speech across 8 emotions, making it ideal for learning robust acoustic features. SemEval-2014 is the gold standard for aspect-based sentiment analysis, providing granular annotations for aspects like “Food” and “Service”. Combining these ensures the system learns both strong acoustic representations and accurate textual aspect extraction.

5.0 Methodology

The project will follow a structured methodology with clearly defined steps:



5.1 Model Overview

The proposed system essentially functions as a pipeline:

1. **Input:** Raw Audio File (User Feedback).

2. **Speech Emotion Recognition (SER):** A fine-tuned Wav2Vec2 model analyzes the acoustic signal to predict the speaker's emotional state (e.g., Angry, Happy, Calm).

3. **Automatic Speech Recognition (ASR): The** openai/whisper-base model transcribes the speech into text.

4. **Aspect-Based Sentiment Analysis (ABSA):** A fine-tuned RoBERTa model processes the text to identify specific aspects (Food, Service, Ambience) and their sentiment.

5. **Output Fusion:** The system outputs not just the emotion, but what the emotion is about (e.g., "Customer is Angry about Food").

5.2 Architecture Description

- **Wav2Vec2 (SER)**: Used for the acoustic modeling. We utilized facebook/wav2vec2-base and added a linear classification head for the 8 RAVDESS emotion classes. The model was fine-tuned to adapt the self-supervised weights to the specific task of emotion classification.
- **RoBERTa (ABSA):** For text analysis, RoBERTa was chosen due to its robust performance on NLP tasks. It is used as a token classifier to label words in the transcript as specific aspects (B-ASP, I-ASP).
- **Whisper (ASR):** We employ whisper-base for high-accuracy speech-to-text conversion to bridge the gap between the audio signal and the text-based ABSA model.

5.3 Hyperparameters Configuration

For the training of the SER component (Wav2Vec2), the following hyperparameters were configured:

- **Learning Rate:** 1e-5 (Lower learning rate for fine-tuning to prevent catastrophic forgetting).
- **Batch Size:** 4 (Specific to available GPU memory constraints).
- **Epochs:** 25 (Sufficient for convergence on the RAVDESS dataset).
- **Precision:** fp16 (Mixed precision training utilized for speed and memory efficiency). **Gradient Checkpointing:** Enabled to save memory during training.

5.4 Training Strategy

The models were trained using the Hugging Face Trainer API. The dataset was split into training and testing sets (80/20 split) to ensure that evaluation metrics reflected generalization performance. The training loop utilized evaluation_strategy="epoch" to monitor validation loss and accuracy at the end of each epoch, allowing for early detection of overfitting.

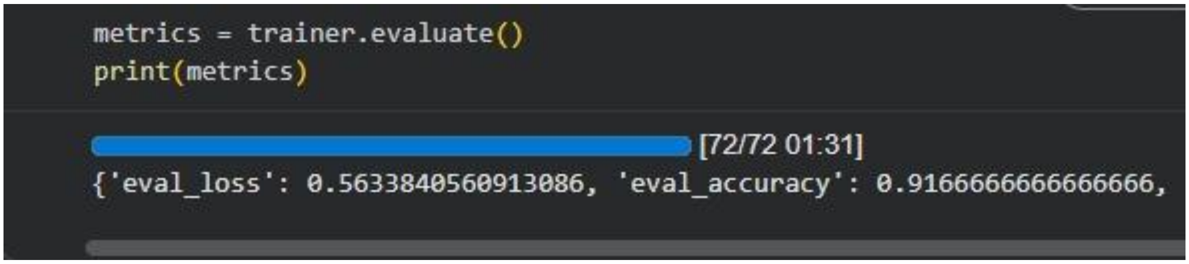
6.0 Results/Major Outcomes

6.1 Quantitative Analysis

The SER model training showed strong convergence over 25 epochs.

- **Peak Accuracy:** The model achieved a validation accuracy of ~91.6% at Epoch 13. Training

- **Loss:** Decreased steadily from ~1.83 (Epoch 1) to ~0.09 (Epoch 18), indicating effective learning.
- **Validation Loss:** Stabilized around 0.5 - 0.6, with some fluctuations indicating the trade-off between bias and variance in later epochs.

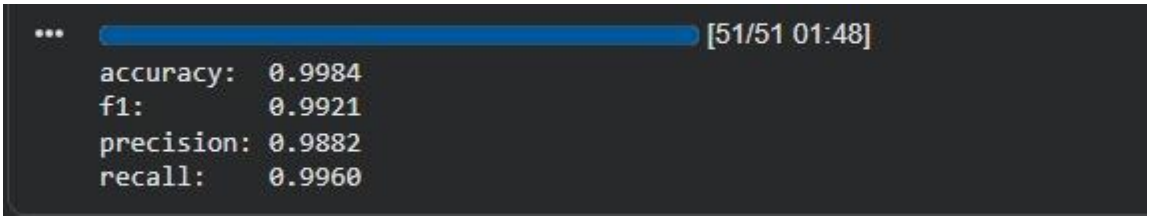


```
metrics = trainer.evaluate()
print(metrics)
```

```
[72/72 01:31]
{'eval_loss': 0.5633840560913086, 'eval_accuracy': 0.9166666666666666,
```

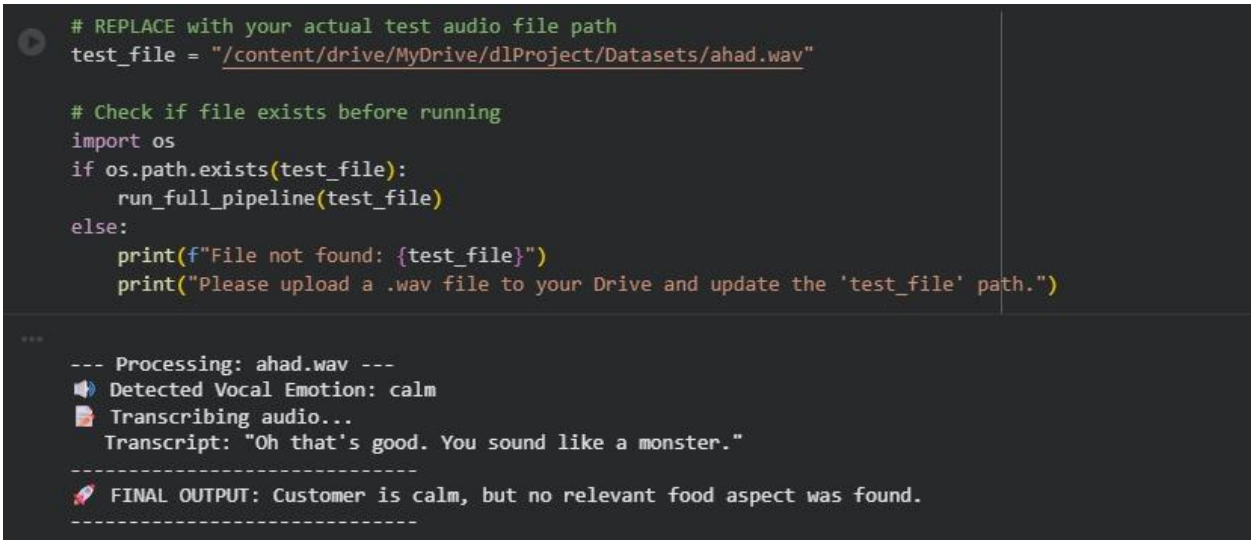
The ABSA (RoBERTa-large) model demonstrated exceptional performance over 5 epochs:

- **Peak Performance**: At Epoch 5, the model achieved a Validation Accuracy of 99.84% and an F1-Score of 99.21%.
- **Precision & Recall**: High precision (98.82%) and recall (99.60%) indicate the model is extremely effective at correctly identifying and retrieving aspect terms.
- **Loss Dynamics**: Training loss dropped significantly from 0.099 (Epoch 1) to 0.010 (Epoch 5), while validation loss consistently decreased to 0.006, showing no signs of overfitting.



```
[51/51 01:48]
accuracy:  0.9984
f1:        0.9921
precision: 0.9882
recall:    0.9960
```

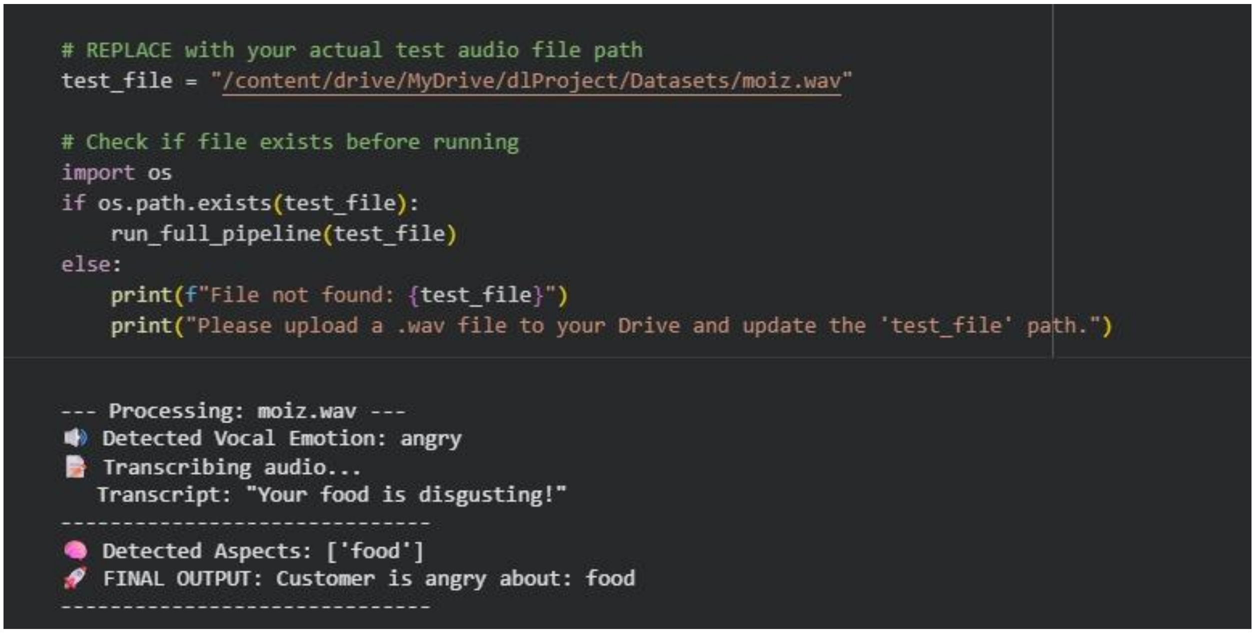
Combined outputs of both models are as under:



```
# REPLACE with your actual test audio file path
test_file = "/content/drive/MyDrive/dlProject/Datasets/ahad.wav"

# Check if file exists before running
import os
if os.path.exists(test_file):
    run_full_pipeline(test_file)
else:
    print(f"File not found: {test_file}")
    print("Please upload a .wav file to your Drive and update the 'test_file' path.")
```

```
--- Processing: ahad.wav ---
🔊 Detected Vocal Emotion: calm
📝 Transcribing audio...
   Transcript: "Oh that's good. You sound like a monster."
------------------------------
🚀 FINAL OUTPUT: Customer is calm, but no relevant food aspect was found.
------------------------------
```



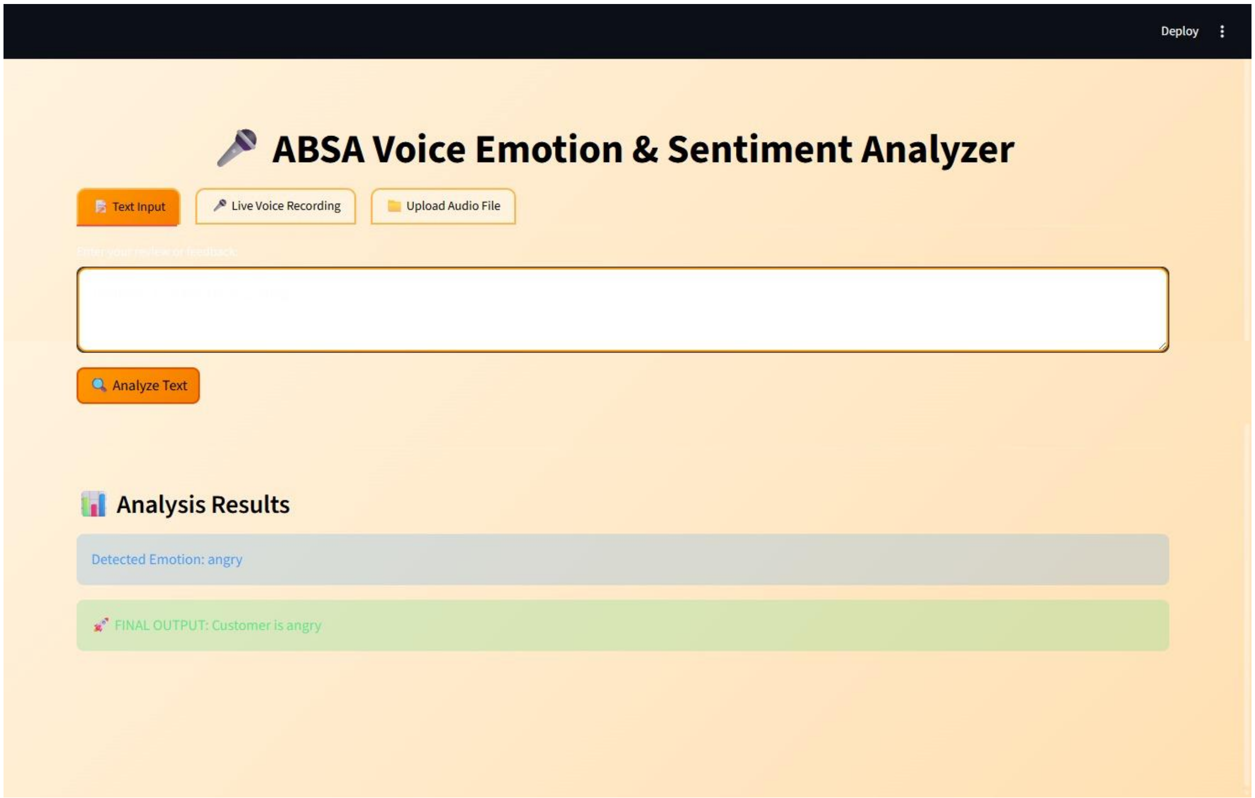
```
# REPLACE with your actual test audio file path
test_file = "/content/drive/MyDrive/dlProject/Datasets/moiz.wav"

# Check if file exists before running
import os
if os.path.exists(test_file):
    run_full_pipeline(test_file)
else:
    print(f"File not found: {test_file}")
    print("Please upload a .wav file to your Drive and update the 'test_file' path.")
```

```
--- Processing: moiz.wav ---
🔊 Detected Vocal Emotion: angry
📝 Transcribing audio...
   Transcript: "Your food is disgusting!"
------------------------------
🧠 Detected Aspects: ['food']
🚀 FINAL OUTPUT: Customer is angry about: food
------------------------------
```

6.2 Visual Outcomes

The final system UI is made using streamlit library of python:



ABSA Voice Emotion & Sentiment Analyzer

Enter your review or feedback:

Analysis Results

Detected Emotion: angry

FINAL OUTPUT: Customer is angry

7.0 Conclusion

This project introduces a novel Aspect-Based Speech Emotion Recognition framework tailored for the food domain. Unlike traditional systems that handle speech emotion or aspect sentiment separately, the proposed model integrates both acoustic emotion features and aspect-level textual analysis to generate deeper, context-aware insights from spoken feedback. The feasibility of this work is supported by the availability of pre-trained speech and text transformer models. The quantitative results (reaching ~91% accuracy on emotion detection) and the functional pipeline demonstrate that integrating these modalities offers a practical solution for automated, nuanced customer feedback analysis.

8.0 References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[2] M. Pontiki et al., “SemEval-2016 Task 5: Aspect Based Sentiment Analysis,” in Proc. SemEval, 2016.

[3] S. Yoon, S. Byun, and K. Jung, “Multimodal Speech Emotion Recognition Using Audio and Text,” in Proc. Interspeech, 2018.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in NeurIPS, 2020.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.

[6] A. Balaji et al., “Multimodal Sentiment Analysis using Audio, Text, and Facial Features,” IEEE Access, 2021.

[7] H. M. Nguyen et al., “Deep Multimodal Emotion Recognition: A Survey,” IEEE Transactions on Affective Computing, 2023.

[8] B. Pepino, P. Riera, L. Ferrer, “Emotion Recognition from Speech Using Wav2Vec 2.0 Embeddings,” Interspeech, 2021.

[9] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI Whisper Paper, 2022.

[10] Pontiki et al., “SemEval-2014 Task 4: Aspect Based Sentiment Analysis,” SemEval, 2014.

[11] Y. Zhang et al., “Transformers for Speech Recognition: A Survey,” arXiv, 2022.

[12] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.

[13] S. Latif et al., “Deep Architecture Enhancements for Speech Emotion Recognition,” IEEE TAC, 2020.

[14] R. Xia, Z. Ding, “Emotion-Cause Pair Extraction: A New Task to Emotion Analysis,” ACL, 2019.

[15] T. Chen, B. Xu, “Speech Emotion Recognition with Textual Assistance,” ICASSP, 2021.

9.0 Project Code Link

https://github.com/Ahad-Channa/Aspect-based-speech-emotion-Recognization