A Comparative Analysis of Text and Speech Modalities for Emotion Detection: Leveraging ML, DL, and Fine-Tuned LLMs and LAMs in a Cultural Niche

BhanuPrakash Yendluri, Dhwani Bhandari, Prasad Somvanshi, Rahul Enamalamanda, Swaijit Singh Sandhu,
Yugandhar Naidu Vankayalapati
Vijaybhoomi Scool of Science and Technology, Vijaybhoomi University, Greater Mumbai, India
Email: {bhanu.prakash, dhwani.bhandari, prasad.balaji, rahul.enamala, swaijit.singh, vankayalapati.yugandhar}
@vijaybhoomi.edu.in

Abstract—**Accurately recognizing emotions in text and speech data remains a challenge due to inherent subjectivity and cultural variations. This paper investigates emotion detection in Indian university students (aged 18-25) utilizing machine learning, deep learning, and Large Language Models (LLMs). We analyze both text and speech data encompassing six basic emotions (anger, disgust, fear, happiness, sadness, and surprise).**

Our research offers several key contributions. First, we conduct a comparative analysis of different model types. This analysis identifies LLMs (particularly Bert_base) as the leader in text emotion detection (F1-score: 0.7528). Large Audio Models (LAMs), specifically facebook/data2vec-audio-base, show promise for speech analysis (F1-score: 0.2824). Second, we incorporate both text and speech data, along with first-person and second-person viewpoints, to gain a more comprehensive understanding of emotional expression across modalities and perspectives. Third, to validate model performance, a separate group of students provided human evaluation (third-person perspective). This human benchmarking resulted in F1-scores of 0.748 for text and 0.789 for speech. Finally, we introduce a stacked model combining the best performing LLM for text and LAM for speech. This stacked model achieves the highest overall F1-score (84%), demonstrating the effectiveness of multimodal approaches for emotion detection.

Our findings highlight the potential of LLMs and LAMs for emotion detection in specific demographics, while also underlining the ongoing challenges in speech analysis. The stacked model's success underscores the efficacy of multimodal approaches. This research lays the groundwork for developing culturally-sensitive and contextually-aware emotion detection models with applications in education, mental health, and personalized online experiences.

I. Introduction

EMOTIONS are complex psychological states encompassing physiological, behavioral, and cognitive changes that significantly influence human experience(1). Recognizing emotions, however, presents a substantial challenge due to their inherent subjectivity. The perceived meaning of an emotional expression can vary depending on the perspective (first-person, second-person, or third-person)(2) and the cultural background of the observer(3).

While significant advancements have been made in developing emotion detection models, including the potential of Large Language Models (LLMs)(4), accurately identifying emotions remains a hurdle, particularly within specific modalities like text(5) and speech. Text lacks the rich vocal cues, such as pitch and tone, present in speech, making it difficult to discern sarcasm or subtle emotional nuances. Conversely, speech itself can be ambiguous due to regional accents or the use of sarcasm(6). These inherent challenges are further amplified when analyzing emotions expressed by a specific demographic group, such as Indian university students (aged 18-25).

This demographic is particularly interesting due to their core social media(7) and digital user status. Accurate emotion detection in this group holds immense practical significance. Businesses can leverage this technology to tailor online experiences and marketing strategies to resonate better with young consumers. Furthermore, in the healthcare domain(8), early detection of potential mental health issues is crucial. Given the concerning rate of suicide among university students in India(9), emotion detection tools could serve as valuable screening mechanisms, enabling healthcare providers to identify students at risk and intervene proactively. Educational institutions could also benefit from this technology by implementing improved support systems tailored to the emotional well-being of their students(10).

In this paper, we present a comparative analysis of emotion detection methods in the context of Indian university students (aged 18-25). Leveraging a dataset comprising text responses and speech recordings from university students, encompassing Ekman's six basic emotions (anger, disgust, fear, happiness, sadness, and surprise)(11).

Our key contributions are:

- **Comparative Analysis:** We conduct a comparative analysis of emotion detection performance using machine learning, deep learning, and LLM-based approaches, specifically tailored to this demographic.
- **Multimodal and Multi-Perspective Analysis:** We incorporate both text and speech data, along with first-person and third-person viewpoints, to gain a more

comprehensive understanding of emotional expression, acknowledging the role of self-reporting and external observations in assessing emotions.
- **Human Benchmarking for Validation:** To ensure objectivity, we introduce human benchmarking(12) with a separate group of students, providing a third-person perspective and validating the models' performance against human judgment.

The findings from this research will contribute valuable insights to the field of emotion detection, paving the way for developing culturally sensitive and contextually aware models. This research has the potential to significantly impact various domains where understanding student emotions is crucial.

In the following sections, we delve into the methodology, results, and implications of our study, paving the way for future research endeavors in the realm of emotion analysis and computational psychology.

II. Interpretation of Emotions

The human experience is deeply intertwined with emotions. These complex psychological states encompass physiological changes, behavioral reactions, and cognitive processes.(13) (14) Recognizing emotions, however, presents a significant challenge due to their inherent subjectivity. The perceived meaning of an emotional expression can vary depending on several factors, including:

- **Perspectives** (1st, 2nd, 3rd person)
- **Medium** (text, speech, image, video)

A. Perspectives

Change in perspective can significantly affect the interpretation of emotions, introducing nuances and biases that influence how emotions are perceived and understood(2). In the context of our study with 18-25-year-old Indian university students, the shift in perspective between first-person, second-person, and third-person observations can impact emotion interpretation in both text and speech data(15).

1) **First-Person Perspective**: In the first-person perspective, individuals directly experience and express their emotions. This perspective offers an intimate insight into the individual's internal state, providing rich and authentic emotional expressions. However, it may also be subject to biases or limitations inherent in self-reporting, such as social desirability bias or difficulty articulating complex emotions.
2) **Second-Person Perspective:** The second-person perspective involves observing and interpreting emotions from the viewpoint of an external observer interacting with the individual. This perspective allows for a more objective assessment of emotions, as it is based on external cues and behaviors. However, interpretations may still be influenced by the observer's own biases or preconceptions about the individual and their emotional expressions.
3) **Third-Person Perspective:** The third-person perspective entails observing and interpreting emotions from a detached standpoint, without direct involvement in the interaction. This perspective offers a broader context for understanding emotions, as it allows for comparisons across different individuals or situations. However, interpretations may be prone to projection or misinterpretation, as the observer relies solely on external cues and contextual information.

B. Medium

*1) **Challenges in Interpreting Emotions from Text Data**:* The collection of text data poses several challenges, particularly in interpreting emotions accurately. Emotions expressed through text are often subtle and context-dependent, presenting difficulties in discernment.(16) This challenge is further compounded by variations in interpretation based on the perspective of the reader(5), especially among 18-25-year-old Indian university students. Their unique cultural backgrounds and digital communication habits influence their perception of emotions conveyed through text. The nature of textual data introduces additional complexities for emotion detection. Emotions are often subtly expressed, and multiple emotions may coexist within a single piece of text. Ambiguities, sarcasm(17), slang, and multilingualism further complicate the interpretation process. These challenges underscore the need for advanced techniques and methodologies in emotion extraction from text.

*2) **Challenges in Interpreting Emotions from Speech Data**:* Similar to text data, interpreting emotions from speech data presents its own set of challenges. Emotions conveyed through speech rely heavily on vocal cues, intonation, and other auditory signals, which can be nuanced and context-dependent(17) (18). Furthermore, the perspective from which emotions are observed—whether second-person or third-person—can influence interpretation, particularly among 18-25-year-old Indian university students.

III. Human Benchmarking

To address the inherent subjectivity of emotion detection and validate our model results, we introduced the concept of human benchmarking in our study. A set of participants who had not been previously exposed to the data were asked to label the data based on their interpretation of it.(12) Providing a third-person perspective for both textual and speech data.

Methodology:

- **Participant Selection:** We recruited a new group of university students (aged 18-25) with no prior knowledge of the data used in the study. This ensured unbiased evaluation and reflected a similar demographic to the original data collection.
- **Assignment Instructions:** Participants received clear instructions outlining the task. These instructions included:
  - A brief explanation of the study's purpose (emotion detection from text and speech data).

  - Definitions of the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise)(11) used in the study.
  - Examples of how each emotion might be expressed in text and speech.
  - The platform they would use to access the data (text excerpts or audio recordings).
- **Evaluation Process:** Participants were presented with anonymized text excerpts and speech recordings from the original data collection. They were instructed to label each data point according to the emotion they perceived to be most dominant.

By incorporating human benchmarking, we aimed to obtain a more holistic perspective on emotion detection in our study. By comparing human interpretations with the results of our models, we highlighted the differences in emotion detection between humans and machines(12) (19), further validating the complexities and nuances involved in emotion detection using computers.

IV. Data Collection

For this study, we opted to utilize Ekman's model of six basic emotions(11) as a foundation for our data collection and building the classifier models. It involves the six basic emotions, namely: happiness, sadness, anger, fear, surprise, and disgust.

A. *Text Data Collection Procedure*

To gather text data, we employed a structured questionnaire designed to elicit responses reflecting Ekman's six basic emotions. We utilized a self-reported emotion (SRE) questionnaire, aligning with established research on emotion detection through SRE surveys.(20) This method yielded a total of 79 responses for each emotion category, providing a valuable dataset for analysis.

B. *Speech Data Collection Procedure*

The speech data collection involved recording responses from university students, specifically chosen to align with the demographic of interest (in alignment with other speech datasets collected like RAVEDESS(21)). Thirteen students, comprising eight males and five females, participated in the data collection process. This approach ensured consistency with the written scenarios while capturing emotional expression through speech.

By employing both text and speech data collection methods, we aim to gain a comprehensive understanding of how emotions are experienced and expressed by young Indian adults.

V. Data Preprocessing

Both text and speech data underwent rigorous cleaning procedures to ensure consistency and quality. All entries with NA values, ”nothing,” and incorrect data such as instances where participants indicated they never felt a particular emotion were removed. After this cleaning process, we were left with 440 valid entries, distributed evenly across all six emotions.

A. *Text Data Preprocessing*

Text data underwent thorough preprocessing steps to eliminate noise, punctuation, and irrelevant information such as emojis. Tokenization and lemmatization techniques were applied to standardize the textual representations. Tokenization involved breaking down the text into smaller units, while lemmatization reduced words to their base or dictionary form.(22)

B. *Speech Data Preprocessing*

For speech data preprocessing, we initiated by applying padding to ensure uniformity across all audio files. Subsequently, we extracted speech features, including Mel Spectrogram and MFCC (Mel-Frequency Cepstral Coefficients)(23). These features provide a comprehensive dataset for training the models built for emotion detection through speech, offering both semantic relationships and low-level features essential for accurate emotion differentiation.

Overall, meticulous cleaning of both text and speech data laid the foundation for accurate and reliable emotion detection analyses.

VI. Models Built for Textual Data

A. *Machine Learning*

For the machine learning approach, we began with feature extraction, converting text data into TF-IDF (Term Frequency-Inverse Document Frequency)(24) vectors using TF-IDF Vectorizer. These preprocessed TF-IDF vectors served as input features for all models. Subsequently, hyperparameter tuning was performed using GridSearchCV across all three models to optimize their performances.

Models Built:

1) **Decision Tree:** Decision Tree is a non-parametric supervised learning method widely used in emotion detection tasks due to its simplicity and interpretability(25)(26). It builds a tree-like model where each node represents a feature (such as words in text data) and each branch represents a decision based on that feature. The best-performing Decision Tree model had a maximum depth of 20, minimum samples per leaf of 1, and minimum samples per split of 10, achieving an accuracy of 0.477.

2) **Random Forest:** Random Forest is an ensemble learning method commonly employed for emotion detection tasks(27). It combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of features and data points, and the final prediction is made by averaging the predictions of all trees. The best-performing Random Forest model had a maximum depth of 20, minimum samples per leaf of 1, minimum samples per split of 2, and 200 estimators (trees), achieving an accuracy of 0.579.

3) **Support Vector Machine (SVM):** Support Vector Machine (SVM) is a powerful classification algorithm widely used in emotion detection due to its ability to find the optimal hyperplane to separate different emotion classes(28)(29). In linear SVM, emotions are separated by a linear decision boundary. The best-performing SVM model had a regularization parameter (C) of 10 and employed a linear kernel, achieving an accuracy of 0.579.

B. Deep Learning

In the deep learning approach, text data underwent tokenization and padding using the Tokenizer from Keras, ensuring consistent sequence lengths across all data samples. Additionally, pre-trained GloVe word embeddings were incorporated to enrich the models' understanding of word semantics. An embedding matrix was generated from the loaded GloVe embeddings and the tokenizer's word index. The padded tokenized sequences, along with the embedding matrix, served as input features for the models. Dropout regularization was applied to prevent overfitting across all models.

Models Built:

1) **RNN (Recurrent Neural Network):** RNNs are a class of neural networks designed to effectively process sequences of data by incorporating feedback loops. In the context of emotion detection, RNNs can capture temporal dependencies in textual or speech data, enabling them to effectively analyze the sequential nature of emotions expressed over time(30). The best-performing RNN model utilized a stack of two SimpleRNN layers with 128 and 64 hidden units respectively, applying dropout regularization to prevent overfitting.

2) **LSTM (Long Short-Term Memory):** LSTMs are a variant of RNNs specifically designed to address the vanishing gradient problem, enabling them to capture long-range dependencies in sequential data. In emotion detection tasks, LSTMs excel at capturing nuanced emotional patterns(31)(32)(33) and long-term contextual information present in textual or speech data. The optimal LSTM configuration in our study comprised two LSTM layers with 128 and 64 hidden units respectively, implementing dropout regularization for enhanced generalization.

3) **BiLSTM (Bidirectional Long Short-Term Memory):** BiLSTMs extend the capabilities of traditional LSTMs by processing input sequences in both forward and backward directions, allowing them to capture dependencies from past and future contexts simultaneously. This bidirectional processing enhances the model's ability to understand complex temporal relationships inherent in emotional expressions(34). The top-performing BiLSTM model in our experiments consisted of two Bidirectional LSTM layers with 128 and 64 hidden units each, utilizing dropout regularization to mitigate overfitting.

C. Large Language Models (LLMs)

Text data underwent tokenization using specific tokenizers associated with each LLM. For Bert_base, the data was tokenized using the Bert tokenizer, while DistilRoberta-base Text data underwent tokenization using specific tokenizers associated with each LLM. For Bert_base, the data was tokenized using the Bert tokenizer, while DistilRoberta-base utilized the Auto tokenizer from transformers. Inputs for these models were converted to tokenized inputs and labels into PyTorch DataLoader format. For Sentence-BERT (SBERT), the 'label' column was one-hot encoded, and the data was converted into dataset.dict format with features including 'text_data', 'Anger', 'Disgust', 'Fear', 'Happiness', 'Sadness', 'Surprise', and 'labels', totaling 329 rows. Additionally, Cohere required at least 40 examples to start fine-tuning, with each label needing a minimum of 5 examples and at least 2 unique labels.

Models Built:

1) **Bert_base:** Bert_base is a transformer-based model fine-tuned for emotion detection tasks(35). It leverages the Bidirectional Encoder Representations from Transformers (BERT) architecture to capture bidirectional context, enabling robust understanding of textual data. The best-performing model was optimized using the AdamW optimizer with a learning rate of 2e-5 and trained for three epochs.

2) **DistilRoberta_base:** DistilRoberta_base is a distilled version of the RoBERTa(36) model, specifically designed for efficient and effective emotion detection. It utilizes the attention mechanism and transformer architecture to process input sequences. The best-performing model employed a batch size of 128, a learning rate of 1.5e-6, and was trained for 40 epochs.

3) **Sentence-BERT (SBERT):** SBERT is a variant of the BERT model fine-tuned for sentence-level tasks such as emotion detection. It utilizes Siamese and triplet networks to learn fixed-size sentence embeddings, allowing for efficient similarity computations(37). The best-performing SBERT model utilized a pre-trained embedding model ("avsolatorio/GIST-small-Embedding-v0") and was trained for 20 iterations using cosine similarity loss.

4) **Cohere:** Cohere(38) provides access to a pre-trained language model called "Embed" specifically designed for tasks like emotion detection. This model leverages advanced natural language processing techniques and a proprietary architecture optimized for understanding nuanced emotional expressions in text data. While specific hyperparameters used for our implementation are not disclosed, the Cohere Embed model achieved an F1-Score of 70.37% in our experiments.

VII. Models Built for Speech Data

A. Machine Learning

For the machine learning approach to speech data analysis, feature extraction is conducted utilizing the MFCC technique, extracting 30 Mel-frequency cepstral coefficients (MFCCs) to capture essential spectral characteristics of the audio signal. This process aids in representing the speech data in a compact and informative manner, facilitating subsequent analysis and classification tasks effectively. Following this, several machine learning models were built and optimized using Grid Search CV to enhance the respective model performances.

Models Built:

1) **Support Vector Machine (SVM):** SVM is a robust classification algorithm applied to various domains, including speech data analysis(39). The best-performing SVM model utilized a regularization parameter (C) of 100, a gamma value of 0.001, and employed an RBF kernel.

2) **Naive Bayes Classifier:** Naive Bayes classifiers(40) are known for their simplicity and efficiency in handling high-dimensional data like speech features. The model was optimized with a var_smoothing parameter of 0.1 through Grid Search CV.

3) **Random Forest Classifier:** Random Forests(41) are ensemble learning methods effective in capturing complex relationships in data. The optimized Random Forest model had a maximum depth of 10, minimum samples per leaf of 4, minimum samples per split of 5, and 100 estimators (trees), achieving an accuracy of 20.3%.

4) **Gradient Boosting Machine (GBM):** GBM is an ensemble learning technique that builds models iteratively, focusing on instances where the previous models performed poorly. The tuned GBM model employed a learning rate of 0.05, a maximum depth of 5, 400 estimators, and a subsample ratio of 0.5.

5) **K-Nearest Neighbors (KNN):** KNN is a simple yet effective algorithm that classifies objects based on the majority vote of their neighbors. The optimized KNN model utilized Euclidean distance as the metric, with 5 nearest neighbors and uniform weighting.

B. Deep Learning

While building Deep Learning models for Emotion Detection via Speech, padding is applied to the audio to standardize the length according to the maximum length requirement, facilitating uniform processing across all audio files. For feature extraction, the MFCC technique is employed with 30 coefficients extracted to capture key spectral characteristics, providing valuable insights into the speech signal's frequency content. Additionally, Mel-Spectrogram features are extracted with 128 Mel-frequency bands to capture the distribution of energy across different frequencies, enhancing the representation of audio signals for subsequent analysis(42).

Models Built:

1) **CNN:** A CNN is a powerful architecture adept at learning spatial features from data. In our case, the CNN extracts relevant features from the Mel-Spectrogram representations of the speech samples (43). This model comprised 4 parallel 2D convolutional and max pooling layers followed by 2 fully connected layers (sizes 400 and 200) with batch normalization. ReLU activation functions were used throughout the network. The final layer feeds into a Softmax output layer, classifying emotions based on the learned features.

2) **CNN+ BiLSTM:** This architecture combines the strengths of CNNs and Long Short-Term Memory (LSTM) networks. The CNN, similar to the previous model, extracts spatial features. LSTMs excel at capturing temporal dependencies within sequences, crucial for understanding the flow of emotions within speech(44). This combined model utilized 2 convolutional and max pooling layers followed by an LSTM layer of size 30 (ReLU activation). The final layer again feeds into a Softmax output layer for emotion classification.

C. Audio Language Models (ALMs)

For building ALM, audio data is initially read using the librosa library, ensuring a consistent sample rate of 16000 Hz or as specified by the model's requirements (instead of 22050 Hz). Following this, feature extraction is performed utilizing the AutoFeatureExtractor from the transformers library. Notably, this feature extraction process is tailored individually for each pre-trained model, ensuring unique and optimized representations suitable for analysing emotions through speech.

To obtain pretrained models, the AutoModelForAudioClassification function from the transformers library is utilized. Furthermore, training arguments are specified using TrainingArguments, encompassing parameters such as evaluation strategy set to "epoch", save strategy configured for "epoch", a total limit of 5 for saving, a learning rate of 3e-5, a training batch size of 32 per device, gradient accumulation steps of 4, an evaluation batch size of 32 per device, 100 training epochs, a warmup ratio of 0.1, logging steps every 10 iterations, and the model is set to load the best model at the end of training based on accuracy metric.

Models Built:

1) **facebook/data2vec-audio-base:** The "facebook/data2vec-audio-base" model(45) is designed with advanced techniques tailored for emotion detection tasks, leveraging its sophisticated architecture to extract intricate patterns and representations from audio data. Its effectiveness lies in its ability to comprehend subtle nuances in vocal expressions and contextual cues

indicative of various emotions.

2) **facebook/wav2vec-base:** The ”facebook/wav2vec-base” model is renowned for its robust performance in speech-related tasks, utilizing self-supervised learning techniques to learn hierarchical representations from raw audio signals. Its adeptness in discerning emotional states from audio inputs is attributed to its comprehensive understanding of speech dynamics and semantics.

3) **microsoft/unispeech-sat-base:** The ”microsoft/unispeech-sat-base” model demonstrates prowess in speech emotion analysis(46), characterized by its capability to handle diverse linguistic features and acoustic variations inherent in speech data. Its reliability for discerning emotional states across different demographic groups and languages makes it a suitable choice for emotion detection tasks.

4) **asapp/sew-d-base-100k:** Conversely, the ”asapp/sew-d-base-100k” model(47) excels in processing speech data with diverse linguistic features and acoustic variations, making it adept at discerning emotional cues across various demographic groups and languages. Its effectiveness lies in its comprehensive understanding of speech dynamics and semantics.

5) **microsoft/wavlm-base-plus:** Lastly, the ”microsoft/wavlm-base-plus” model stands out for its innovative approach to audio language modeling, incorporating advanced mechanisms for capturing temporal dependencies and semantic relationships in speech signals. Its sophisticated architecture enables accurate analysis of emotional states expressed through speech, ensuring robust performance in emotion detection tasks.

VIII. Results

This section delves into the outcomes of our diverse text and speech models utilized to categorize the six fundamental emotions on a dataset specifically targeting Indian university students (aged 18-25). A crucial aspect of this analysis involves selecting an appropriate metric to evaluate the performance of our models. Since emotion detection constitutes a multi-class classification problem, where each data point belongs to one of several distinct categories (e.g., happy, sad, angry), traditional metrics like accuracy can be misleading. Accuracy simply measures the proportion of correctly classified instances, but it doesn't consider the distribution of these classes.

In multi-class classification problems, particularly those with imbalanced datasets, a model might achieve high accuracy by simply predicting the majority class. This wouldn't necessarily reflect a true understanding of the emotions present in the data. Therefore, for a more comprehensive evaluation, we have chosen F1-score as the primary metric.

F1-score incorporates both precision and recall, providing a balanced view of a model's performance. Precision measures the proportion of predicted positive labels that are actually correct, while recall reflects the model's ability to identify all true positive instances. F1-score takes the harmonic mean of these two metrics, offering a single score that penalizes models that excel in either precision or recall at the expense of the other. By utilizing F1-score, we gain a more nuanced understanding of how well our models are classifying emotions across all six categories. Tables 1 and 2 present a comprehensive overview of the F1-scores achieved by all models within the text and speech domains, respectively.

The table I and II below summarize the results of all models across all categories.

A. *Text Emotion Detection*

Table 1 summarizes the performance of all models across categories(Machine Learning, Deep Learning, Large Language Models) for emotion detection through the medium, text.

In analyzing the results, it is evident that Bert_base model emerged as the top-performing model with an F1-Score of 0.7528, showcasing its remarkable capability in accurately detecting emotions from textual data. This reinforces the groundbreaking advancements in emotion detection, particularly with the introduction of Large Language Models (LLMs), even other LLMs show better performance compared to the traditional ML and DL model, despite the limited size of the data which also plays a crucial role in the performance of these advanced Language Models-LLMs . LLMs have revolutionized the field by enabling models to grasp complex linguistic nuances and semantic representations, contributing to significant improvements in accuracy. Our study, focused on a specific demographic and age group with limited data, demonstrates the immense scope for improvement in emotion detection through text despite its inherent challenges and previous lack of notable results. This underscores the evolving landscape of emotion detection problems and the potential for rapid advancements, as evidenced by other recent research endeavors achieving remarkable accuracies in similar domains.

B. *Speech Emotion Detection*

Similar to the text analysis, we evaluated the performance of all models across categories (Machine Learning, Deep Learning, Large Audio Models) for emotion detection through the medium, speech.

Unlike text analysis, where Sentence-BERT emerged as the clear leader (F1-score: 0.7037), speech emotion detection presented a more challenging landscape. Here, facebook/data2vec-audio-base, a Large Audio Model (LAM), achieved the highest F1-score (0.2824) among the tested models. While this performance is promising, it falls short compared to the results obtained in text analysis.

TABLE I
MODEL PERFORMANCE COMPARISON FOR TEXT

Category	Model	Best Parameters	F1-Score
Machine Learning	Decision Tree	’max_depth’: 20, ’min_samples_leaf’: 1, ’min_samples_split’: 10	0.511
	Random Forest	’max_depth’: 20, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 200	0.58
	SVM	’C’: 10, ’kernel’: ’linear’	0.576
Deep Learning	RNN	’max_words’: 100000, ’maxlen’: 150, ’embedding_dim’: 100	0.354
	LSTM	’max_words’: 100000, ’maxlen’: 150, ’embedding_dim’: 100	0.424
	BiLSTM	’max_words’: 100000, ’maxlen’: 150, ’embedding_dim’: 100	0.493
Large Language Models (LLMs)	Bert_base	max_length=128, optimizer: AdamW (lr=2e-5), epochs=3	0.7528
	DistilRoberta-base	model_name: ’distilroberta-base’, batch_size: 128, lr: 1.5e-6, epochs=40	0.5934
	Sentence-BERT (SBERT)	model_id: ”avsolatorio/GIST-small-Embedding-v0”, loss: CosineSimilarityLoss, iterations=20	0.645
	Cohere - embed-english-v2.0	Dimensions - 4096, Max tokens - 512, Similarity Metric - Cosine Similarity	0.7037

TABLE II
MODEL PERFORMANCE COMPARISON FOR SPEECH

Category	Model	Parameters	F1-Score
Machine Learning	SVM	{’C’: 100, ’gamma’: 0.001, ’kernel’: ’rbf’}	0.2157
	Naive Bayes	var_smoothing parameter: 0.1	0.1846
	Random Forest	{’max_depth’: 10, ’min_samples_leaf’: 4, ’min_samples_split’: 5, ’n_estimators’: 100}	0.2129
	GBM	{’learning_rate’: 0.05, ’max_depth’: 5, ’n_estimators’: 400, ’subsample’: 0.5}	0.233
Deep Learning	CNN	4 parallel 2D convolutional and max pooling layers, followed by 2 fully connected layers (sizes 400 and 200) with batch normalization. ReLU activation throughout. Softmax output layer.	0.2656
	LSTM + CNN	2 convolutional and max pooling layers, followed by an LSTM layer of size 30. ReLU activation throughout. Softmax output layer.	0.2487
Large Audio Models (LAMs)	facebook/data2vec-audio-base	Get pretrained models using AutoModelForAudioClassification from transformers library with specific training arguments.	0.2824
	facebook/wav2vec-base	Same training arguments as data2vec-audio-base.	0.2117
	microsoft/unispeech-sat-base	Same training arguments as data2vec-audio-base.	0.2746
	asapp/sew-d-base-100k	Same training arguments as data2vec-audio-base.	0.2230
	microsoft/wavlm-base-plus	Same training arguments as data2vec-audio-base.	0.211

Several factors might contribute to this disparity. Firstly, the text data likely captured a more direct expression of emotions since participants themself typed their emotional experiences. Speech recordings, on the other hand, were captured by a different person, introducing a layer of third-person perspective that might have diluted the emotional intensity. Additionally, the limited data available for speech analysis compared to text could have hampered the models’ ability to learn complex relationships between speech features and emotions.

Furthermore, speech is inherently more intricate than text for machines to interpret. Vocal nuances, tone variations, and background noise can all influence the emotional content of speech, making it a more challenging signal to decode compared to written words. This highlights the complex nature of human emotion and the difficulty of replicating human-like emotional intelligence in machines.

The observed difference in performance across modalities (text vs. speech) underscores the challenges computers face in understanding, comprehending, and classifying emotions. Human emotion detection is subjective even for other humans. When translated into the realm of machines, which are essentially complex black boxes, the challenge becomes even more significant. Ultimately, it is data that shapes a machine’s ability to detect emotions. With limited data and the inherent complexities of speech, accurate emotion detection in speech remains an ongoing area of research.

Despite the current limitations, the promising performance of LAMs like facebook/data2vec-audio-base suggests a path forward. Further research with larger and more diverse speech datasets, combined with the exploration of multimodal approaches utilizing both text and speech data, holds the potential to bridge the gap between text and speech emotion detection performance.

IX. Comparative Analysis: Human Benchmarking and Top Performers

This section delves into a comparative analysis of the top-performing models across text and speech emotion detection tasks, along with comparisons to human benchmarking results(48) (49). We also introduce a stacked model combining the best performing models from each modality, achieving the highest overall F1-score.

Table III: Top 3 Performing Models and Human Benchmarking

A. *Human Benchmarking*

The human benchmark scores provide a valuable reference point for evaluating the models' performance. As observed in Table 3, human agreement on emotion labeling (F1-score of 0.748 for text and 0.789 for speech) establishes a baseline for the models to strive towards. While both LLMs (Bert_base and Cohere) in text analysis approach human-level performance, there remains room for improvement. Speech emotion detection, on the other hand, presents a more significant challenge for the models, with the best performing LAM (data2vec-audio-base) achieving an F1-score of 0.2824, which is considerably lower than the human benchmark.

B. *Top Performing Models*

Text:

1) Bert_base (F1-score: 0.7528): This LLM emerged as the leader in text emotion detection, demonstrating the effectiveness of LLMs in capturing emotions from written language.
2) Cohere - embed-english-v2.0 (F1-score: 0.7037): This commercially available LLM achieved impressive results, highlighting the potential of pre-trained models for emotion detection tasks.
3) SBERT - setfit (F1-score: 0.6450): This SBERT model showcases the capability of sentence embeddings for emotion classification, offering an alternative approach to LLMs.

Speech:

1) facebook/data2vec-audio-base (F1-score: 0.2824): This LAM achieved the highest F1-score among speech models, paving the way for further exploration of LAMs in speech emotion detection.
2) wav2vec2-base-960h (F1-score: 0.3096): This LAM demonstrates competitive performance, suggesting the potential of different architectures within the LAM category.
3) microsoft/unispeech-sat-base (F1-score: 0.2746): This LAM showcases the promise of pre-trained models for speech emotion analysis.

C. *Stacking*

To leverage the strengths of both modalities, we built a stacked model combining the best-performing models: Bert_base for text and data2vec-audio-base for speech. The stacked model takes the probability outputs for each emotion class from both individual models (12 features in total) and feeds them into an Artificial Neural Network (ANN) with two fully connected layers for further processing. Finally, a Softmax output layer predicts the most likely emotion class. This stacked model achieved an F1-score of 84.09% on our dataset, surpassing the performance of all individual models and demonstrating the effectiveness of multimodal emotion detection.

The confusion matrix for the stacked model is shown below in Figure 1 and Figure 2, demonstrating the model's performance across different emotion classes.

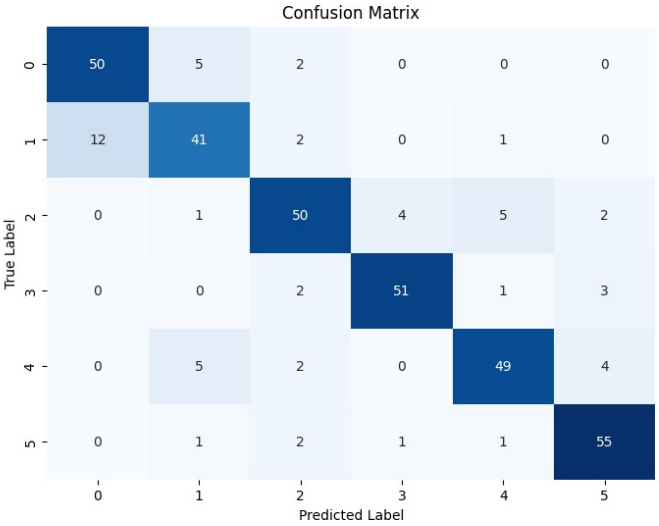

Fig. 1. Confusion Matrix (Test Data)

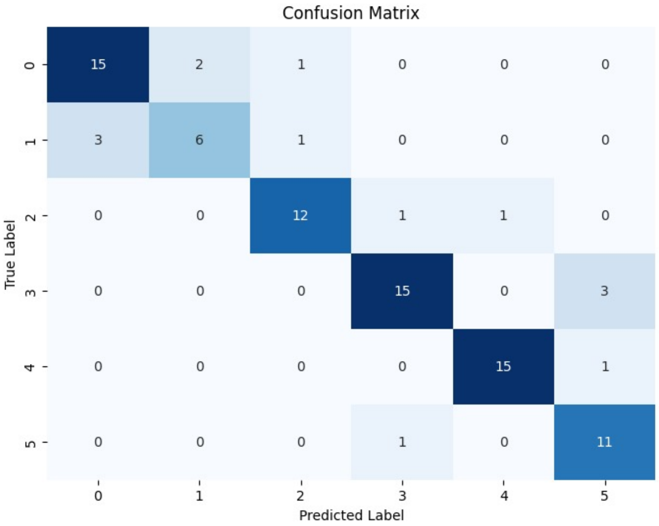

Fig. 2. Confusion Matrix (Test Data)

TABLE III
COMPARATIVE ANALYSIS

Modality	Model Name	Result (F1 Score)	Human Benchmark
Text	Bert_base	0.7528	0.748
	Cohere - embed-english-v2.0	0.7037	
	SBERT - setfit	0.645	
Speech	data2vec-audio-base	0.3324	0.789
	wav2vec2-base-960h	0.3096	
	Unispeech-sat-base	0.2746	
Stacking	Bert_base + data2vec-audio-base	0.8409	

X. CONCLUSIONS

In this work, we embarked on a journey to explore the multifaceted nature of emotion perception and detection between humans and machines. We investigated the potential of deep learning models to analyze not only the semantic content of text (what is being said) but also the underlying emotional cues embedded within speech patterns (how it is being said). Our research focused on exploring these modalities (text and speech) from various perspectives (1st person and 3rd person) while leveraging the power of advanced Language Models (LLMs) and a stacking approach to combine speech and text features.

Our investigation yielded promising results. The deep learning models, particularly the one focused on text emotion detection, effectively learned to identify sentiment and emotional undertones by analyzing factors like word choice, punctuation, and sentence structure. Analyzing speech, however, presented a different challenge. While the models learned to recognize some emotional cues from vocal characteristics like pitch, intonation, and speech rhythm (captured through MFCC features and mel spectrograms), their performance was lower than expected. We hypothesize that this may be due to limitations in the speech data collection process, which involved recordings from actors portraying emotions rather than genuine experiences. This highlights the importance of using datasets that capture authentic emotional expression.

Despite the challenges with speech data, our multifaceted approach yielded significant results. The stacked model combining features from both text and speech achieved a best-performing F1-score of 84%, at par with top performing models in this domain. This achievement signifies the effectiveness of our approach, particularly compared to existing works that utilize multimodal datasets like RAVDESS. While RAVDESS(21) incorporates speech, image, and a combination of speech and image information, our focus on just text and speech, coupled with the analysis of different perspectives, demonstrates the potential of this combined approach. The high F1-score suggests that our model can accurately detect emotions from these modalities, opening doors for various applications.

Furthermore, this achievement is noteworthy considering the data used. While the model was trained on a dataset specifically focused on Indian University students aged 18-25, it still achieved a high F1-score. This demonstrates the power of LLMs and the chosen methods for emotion detection and understanding, even with limited and niche data. This opens exciting avenues for future work, suggesting the possibility of tailoring emotion detection systems to specific target audiences across various domains. By focusing on relevant and authentic data, leveraging these techniques, and delving deeper into capturing nuanced emotions, we can build human-like systems that cater to diverse emotional experiences.

The inclusion of human benchmarking during our study proved invaluable. By incorporating human evaluation alongside machine learning models, we gained crucial insights into the real-world complexities of emotion perception. Human raters, with their inherent understanding of emotional nuances and contextual factors, provided a valuable benchmark for the models' performance. This comparison highlighted the remaining challenges associated with replicating human-like emotional intelligence in machines, underlining the subjectivity and context-dependent nature of emotions.

Overall, our findings contribute significantly to the field of affective computing. By demonstrating the effectiveness of deep learning models for both text and speech emotion detection, particularly when combined through a stacking approach, we pave the way for further advancements in human-computer interaction. This research opens doors to a future where technology can not only understand the information we convey but also the emotions that lie beneath the surface, fostering richer and more emotionally intelligent interactions between humans and machines.

XI. LIMITATIONS AND DIRECTION FOR FUTURE WORK

Our study has yielded promising results with the stacked model achieving a high F1-score. However, we acknowledge several limitations that pave the way for exciting future work:

Data Size and Demographics:

- **Limited dataset size:** The current datasets used for training may limit the generalizability of the models. Expanding the data volume would enhance the models' ability to handle unseen data with different characteristics.
- **Demographic focus:** The current data focuses on a specific demographic (Indian University students aged

18-25). To understand the model's effectiveness across diverse populations, validation on datasets with different demographics is crucial. Emotions can be culturally and age-dependent, necessitating a broader representation in training data.

Multimodal Emotion Detection:

- **Focus on individual modalities:** While stacking performs well, future work could explore training models on datasets that include both text and speech data simultaneously. This ”multimodal” approach would leverage the strengths of each modality for a more comprehensive understanding of emotions. Additionally, collecting data in the first-person perspective for all three modalities (text, speech, image) could provide richer information for model training.

Detailed Emotion Analysis:

- **Limited breakdown:** Currently, we lack a detailed breakdown of emotion-specific accuracies and F1-scores. This analysis is crucial to understand how effectively the model detects each basic emotion. It would enable tailoring solutions to specific target demographics and gaining a deeper understanding of their emotional states.

Capturing Nuanced Emotions:

- **Complexity of human emotions:** Human emotions are multifaceted, influenced by perspective (POV) and the possibility of experiencing multiple emotions simultaneously. Future work could explore advanced techniques for discerning these nuanced emotional states and their interplay.

In conclusion, this research serves as a springboard for further exploration in the realm of emotion detection. By addressing the limitations identified and pursuing the exciting avenues outlined above, we can continue to refine and advance the capabilities of deep learning models in recognizing and understanding human emotions.

XII. Acknowledgement

We would like to express our sincere gratitude to the following individuals and organizations whose invaluable support played a crucial role in the successful completion of this research:

- We are deeply grateful to our esteemed supervisor, Professor Saunak Dutta and Sai Teja Macherla, for their expert guidance, insightful feedback, and unwavering encouragement throughout this project. Their dedication to our academic development and their expertise in the field of emotion detection were instrumental in shaping this research.
- Finally, we express our heartfelt appreciation to all the participants who generously volunteered their time and effort for the human benchmarking process. Their participation was essential in validating our findings and providing a valuable human perspective on emotion detection.

References

[1] J. J. Gross *et al.*, “Emotion regulation: Conceptual and empirical foundations,” *Handbook of emotion regulation*, vol. 2, pp. 3–20, 2014.

[2] J. A. Deonna, “Emotion, perception and perspective,” *dialectica*, vol. 60, no. 1, pp. 29–46, 2006.

[3] B. MESQUITA^1, N. H. Frijda, and K. R. Scherer, “Culture and emotion,” *Handbook of cross-cultural psychology: Basic processes and human development*, vol. 2, p. 255, 1997.

[4] B. V. Kok-Shun, J. Chan, D. Sundaram, and G. Peko, “Intertwining two artificial minds: Chaining gpt and roberta for emotion detection,” in *10th IEEE Asia Pacific Conference on Computer Science and Data Engineering*, 2023.

[5] N. Alvarez-Gonzalez, A. Kaltenbrunner, and V. Gómez, “Uncovering the limits of text-based emotion detection,” *arXiv preprint arXiv:2109.01900*, 2021.

[6] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, “Emotion detection from text and speech: a survey,” *Social Network Analysis and Mining*, vol. 8, no. 1, p. 28, 2018.

[7] B. Gaind, V. Syal, and S. Padgalwar, “Emotion detection and analysis on social media,” *arXiv preprint arXiv:1901.08458*, 2019.

[8] A. H. Saffar, T. K. Mann, and B. Ofoghi, “Textual emotion detection in health: Advances and applications,” *Journal of Biomedical Informatics*, vol. 137, p. 104258, 2023.

[9] V. Patel, C. Ramasundarahettige, L. Vijayakumar, J. Thakur, V. Gajalakshmi, G. Gururaj, W. Suraweera, and P. Jha, “Suicide mortality in india: a nationally representative survey,” *The lancet*, vol. 379, no. 9834, pp. 2343–2351, 2012.

[10] D. T. van der Haar, “Student emotion recognition in computer science education: A blessing or curse?,” in *Learning and Collaboration Technologies. Designing Learning Experiences: 6th International Conference, LCT 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part I 21*, pp. 301–311, Springer, 2019.

[11] P. Ekman, “Are there basic emotions?,” 1992.

[12] C. N. van der Wal and W. Kowalczyk, “Detecting changing emotions in human speech by machine and humans,” *Applied intelligence*, vol. 39, pp. 675–691, 2013.

[13] M. A. Jalal, R. Milner, and T. Hain, “Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition,” in *Proceedings of Interspeech 2020*, pp. 4113–4117, International Speech Communication Association (ISCA), 2020.

[14] S. Aman and S. Szpakowicz, “Identifying expressions of emotion in text,” in *International Conference on Text, Speech and Dialogue*, pp. 196–205, Springer, 2007.

[15] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[16] V. Tripathi, A. Joshi, and P. Bhattacharyya, "Emotion analysis from text: A survey," *Center for Indian Language Technology Surveys*, vol. 11, no. 8, pp. 66–69.

[17] M. Sykora, S. Elayan, and T. W. Jackson, "A qualitative analysis of sarcasm, irony and related# hashtags on twitter," *Big Data & Society*, vol. 7, no. 2, p. 2053951720972735, 2020.

[18] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language—a survey," *Emotion Recognition: A Pattern Analysis Approach*, pp. 237–267, 2015.

[19] P. Damacharla, A. Y. Javaid, J. J. Gallimore, and V. K. Devabhaktuni, "Common metrics to benchmark human-machine teams (hmt): A review," *IEEE Access*, vol. 6, pp. 38637–38655, 2018.

[20] E. Lavoué, G. Molinari, and M. Trannois, "Emotional data collection using self-reporting tools in distance learning courses," in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, pp. 377–378, IEEE, 2017.

[21] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[22] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PloS one*, vol. 15, no. 5, p. e0232525, 2020.

[23] S. Mohmmad and S. K. Sanampudi, "Tree cutting sound detection using deep learning techniques based on mel spectrogram and mfcc features," in *Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022*, pp. 497–512, Springer, 2023.

[24] V. Sundaram, S. Ahmed, S. A. Muqtadeer, and R. R. Reddy, "Emotion analysis in text using tf-idf," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 292–297, IEEE, 2021.

[25] S. Sriram and X. Yuan, "An enhanced approach for classifying emotions using customized decision tree algorithm," in *2012 Proceedings of IEEE Southeastcon*, pp. 1–6, IEEE, 2012.

[26] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[27] P. Vora, M. Khara, and K. Kelkar, "Classification of tweets based on emotions using word embedding and random forest classifiers," *International Journal of Computer Applications*, vol. 178, no. 3, pp. 1–7, 2017.

[28] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021.

[29] S. A. Salam and R. Gupta, "Emotion detection and recognition from text using machine learning," *Int. J. Comput. Sci. Eng*, vol. 6, no. 6, pp. 341–345, 2018.

[30] M. A. Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 718–728, 2017.

[31] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, "Lstm-based text emotion recognition using semantic and emotional word vectors," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6, IEEE, 2018.

[32] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi, and S. K. Shahzad, "Emotion detection of contextual text using deep learning," in *2020 4th International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pp. 1–5, IEEE, 2020.

[33] M. A. Riza and N. Charibaldi, "Emotion detection in twitter social media using long short-term memory (lstm) and fast text," *Int. J. Artif. Intell. Robot*, vol. 3, no. 1, pp. 15–26, 2021.

[34] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE access*, vol. 7, pp. 111866–111878, 2019.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[36] R. Kamath, A. Ghoshal, S. Eswaran, and P. Honnavalli, "An enhanced context-based emotion detection model using roberta," in *2022 IEEE international conference on electronics, computing and communication technologies (CONECCT)*, pp. 1–6, IEEE, 2022.

[37] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," *arXiv preprint arXiv:2209.11055*, 2022.

[38] "Comprehensive analysis: Cohere vs. gpt-4."

[39] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu, *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[40] S. K. Bhakre and A. Bang, "Emotion recognition on the basis of audio signal using naive bayes classifier," in *2016 International conference on advances in computing, communications and informatics (ICACCI)*, pp. 2363–2367, IEEE, 2016.

[41] S. Yan, L. Ye, S. Han, T. Han, Y. Li, and E. Alasaarela, "Speech interactive emotion recognition system based on random forest," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 1458–1462, 2020.

[42] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," *arXiv preprint arXiv:1906.05681*, 2019.

[43] A. Christy, S. Vaithyasubramanian, A. Jesudoss, and M. A. Praveena, "Multimodal speech emotion recognition and classification using convolutional neural network techniques," *International Journal of Speech Technology*,

vol. 23, no. 2, pp. 381–388, 2020.

[44] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[45] Z. Ma, W. Wu, Z. Zheng, Y. Guo, Q. Chen, S. Zhang, and X. Chen, "Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition," *arXiv preprint arXiv:2309.10294*, 2023.

[46] L. Gómez-Zaragozá, Ó. Valls, R. del Amor, M. J. Castro-Bleda, V. Naranjo, M. A. Raya, and J. Marín-Morales, "Speech emotion recognition from voice messages recorded in the wild," *arXiv preprint arXiv:2403.02167*, 2024.

[47] E. Tzagkarakis, "Emotion recognition on scenes of films based on the speech and the image," Master's thesis, $\Pi\alpha\nu\varepsilon\pi\iota\sigma\tau\acute{\eta}\mu\iota$ο $\Pi\varepsilon\iota\rho\alpha\iota\acute{\omega}\varsigma$, 2023.

[48] M. de Velasco, R. Justo, J. Antón, M. Carrilero, and M. I. Torres, "Emotion detection from speech and text.," in *IberSPEECH*, pp. 68–71.

[49] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 112–118, 2018.