International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# Decoding Human Emotions Through Multimodal Analysis Using Deep Learning

**Shanthini C[1], Sukesh N D[2], Arvinth U S[3], Satheesh N P[4]**

[1]*Student, Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam,*
[2]*Student, Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam,*
[3]*Student, Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam,*
[4] *Assistant Professor, Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam.*

---

## ABSTRACT

In the evolving landscape of emotion recognition, this study addresses the inherent shortcomings in existing technology. Recognizing the imperative role emotions play in human interaction, the need for a more nuanced and accurate multimodal emotion recognition system is evident. This research aims to refine multimodal emotion recognition by addressing existing challenges. The problem statement revolves around enhancing the accuracy and depth of emotion recognition through innovative methodologies. Leveraging Convolutional Neural Networks (CNN), Attention layers, and Bi-LSTM networks, the study utilizes a comprehensive approach. Data fusion involves combining audio extracted from video data, followed by fusion combinations of audio and text. The methodology incorporates SoftMax classifiers for feature recognition. The fusion of audio, text, and video data, coupled with the innovative use of CNN-LSTM and Attention layer networks, contributes to this success. The discussion interprets these results, highlighting the efficacy of the proposed approach in addressing the challenges of multimodal emotion recognition. The outcomes signify a substantial advancement in the field, providing a more comprehensive understanding of emotional expressions.

## 1. INTRODUCTION

Understanding emotions is crucial in both artificial intelligence and human-computer interaction, marking a significant advancement in developing empathetic systems. The traditional single-mode approach has limitations in capturing the intricate nuances of human emotions when navigating new technological landscapes. This method primarily focuses on individual modalities such as facial expressions or vocal tones, often overlooking the richness and complexity inherent in emotional experiences. Consequently, there's been a shift towards more sophisticated frameworks capable of seamlessly interpreting and integrating multiple modalities simultaneously. As technology evolves, there's a growing demand for systems that can adeptly interpret various channels of human expression. Recognizing that emotions manifest in diverse ways—including physiological signals, vocal nuances, and facial cues—there's a need for comprehensive methodologies that transcend the constraints of single-mode approaches. Understanding that combining information from different sources is essential for a deeper comprehension of human emotions is propelling the transition towards multimodal systems.

To remain abreast of the swiftly evolving technological landscape, systems must possess the capacity to discern and empathetically respond to emotions. Typical unimodal approaches often fall short when it comes to crafting truly responsive systems that mirror the

intricacies of human emotional experiences. The imperative for technology to transcend mere transactional functionality propels the advancement of intricate multimodal frameworks capable of perceiving and reacting to the entire spectrum of human emotions. Fundamentally, the trajectory of technological advancement aligns with society's aspiration for emotionally astute systems that enhance human-machine interactions. A notable transition in the evolution of intelligent systems is evident in the shift from unimodal to multimodal approaches for emotion detection. This transition reflects a growing recognition of the constraints of conventional methods and the necessity for a more comprehensive, refined approach to facilitate authentic and responsive human-computer interactions. As technology progresses, the path towards empathetic, intelligent systems adept at unraveling the complex tapestry of human emotions continues to unfold.

## 2. LITERATURE SURVEY

Thanveer Shaik (2023) - This paper provides an extensive overview of sentiment analysis techniques in the field of education. Examining lexicon-based, corpus-based, and sophisticated methods such as machine learning and deep learning, the paper comprises case examples involving deep learning models, decision trees, logistic regression, and various degrees of sentiment analysis. An extensive review of sentiment analysis in an educational setting is given in this work.

James Mutinda et al. (2023) - The study used the LeBERT model to conduct experiments on reviews from Yelp, IMDB, and Amazon. LeBERT employed a Convolutional Neural Network (CNN) for classification and BERT word embeddings, sentiment lexicons, and N-grams for text vectorization. Tokenization and N-grams creation were part of the preprocessing. With an accuracy of 82.4, the model was assessed on the remaining 20% of the dataset after being trained on 80% of it.

M. Kalpana Chowdary, et al. (2021) - Using the CK+ dataset, which contains 918 photos and seven emotions, the study used pre-trained CNNs (MobileNet, Resnet50, VGG19, Inception V3) for facial emotion recognition. Utilizing ImageNet's initialized learning weights, modify models to identify emotions. Max pooling, fully linked, and convolutional layers were described in detail in the architectures. Images were resized, layers were frozen, and only completely connected layers were trained in the experiments.

Chen Zeng Zhao et al. (2022) - Deep learning approaches were employed in the study for Speech Emotion Recognition (SER), with open SMILE being utilized to extract features from spectrograms, frame-level features, and utterance-level deep features. An attention-based aggregation features fusion model called AMSNet was presented. The evaluation was conducted using the IEMOCAP and EmoDB datasets.

Ziyang Ma, et al. (2023) - The goal of the project was to employ cutting-edge techniques to advance Speech Emotion Recognition (SER). Using the IEMOCAP dataset, GPT-4 produced text that was emotionally consistent, while Azure Emotional TTS was used to synthesize voice. Features were retrieved using Data2vec, and methodologies such as cross-entropy loss, curriculum learning, and transfer learning were used. In an extensive experimental setup, the framework combined state-of-the-art approaches, software (GPT-4, Azure Emotional TTS), and strategies to advance SER.

Seyed Sadegh Hosseini, et al (2023) - The study used deep learning techniques (Inception-ResNet-v2, CNN-LSTM, and Bi-LSTM) to extract features from text, audio, and video data. Using SoftMax models, data fusion and classification were carried out. The performance evaluation of the suggested multimodal emotion recognition model was conducted using the IEMOCAP dataset.

Zhen-Tao Liu and et al (2023) introduced a new technique for audio emotion identification, employing a hierarchical deep neural network (DNN) model. This model consists of two layers: a high-level emotion classification layer and a low-level feature extraction layer. The latter includes a long short-term memory (LSTM) network to capture temporal dynamics in speech signals, along with a convolutional neural network (CNN) for extracting local features from speech spectrograms. Emotion classification is carried out by a SoftMax layer, while the features extracted by the low-level layer are projected into a high-dimensional space using a fully connected layer. Evaluation of the proposed model on two datasets, AVEC2017 and IEMOCAP, demonstrates its cutting-edge performance across both datasets.

In Hong-Wei Ng et al.'s (2023) study, deep Convolutional Neural Network (CNN) architectures were used in conjunction with transfer learning to recognize emotions. Two distinct deep CNN architectures were pre-trained on the ImageNet dataset as part of the approaches, and then there was a two-stage supervised fine-tuning procedure. The FER-2013 face expression dataset was used for fine-tuning in the first stage, while the EmotiW dataset's features were taken into account in the second stage for adjusting the network weights. Face detection was done using the Viola & Jones face detector in OpenCV. OpenCV was one of the software tools for face identification, and deep learning frameworks were used for CNN training and optimization... The algorithms included Viola & Jones face identification, supervised fine-tuning, and transfer learning. The methods included feature extraction and extensive CNN architecture customization for small dataset emotion recognition. All things considered, the study combined deep learning techniques, datasets, and pre-existing tools to tackle the particular difficulties associated with emotion recognition in real-world situations.

## 3. METHODOLOGY

### A) DATASET COLLECTION:

The MELD (Multimodal EmotionLines Dataset) is a large-scale dataset developed for multimodal emotion analysis research. The MELD dataset has three noteworthy properties.

Multimodal Nature: Because MELD incorporates both textual and audiovisual data, it is a helpful tool for studying multimodal emotion recognition. The collection includes video and audio clips that sync with dialogue extracted from popular television shows. Due to its multimodal nature, researchers can look at the integration of spoken and nonverbal cues to gain a deeper understanding of human emotions.

Diversity of Emotions: MELD includes a broad range of feelings, including surprise, fear, anger, sadness, joy, and contempt, that are expressed in various contexts. The collection captures these emotions in conversational contexts in their organic and spontaneous presentations, offering rich and varied examples for emotion analysis research.

Annotation: MELD manually annotates emotional expressions to offer ground truth labels for each sentence in the dataset. Annotation of emotions at both utterance and dialog levels enables a fine-grained analysis of the emotional dynamics in talks. This annotation technique ensures consistency and quality of emotional labels, which facilitates supervised learning tasks such as emotion categorization and sentiment analysis.

### B) DATA EXTRACTION:

To extract audio data from MP4 video files, we used a video-to-audio extraction procedure in our project. Using the MoviePy library, each video file was loaded as part of a methodical iteration over the video files kept in a specified directory. The audio portion of every video was then extracted and stored in a designated directory as a WAV file. We included strong exception handling procedures to handle possible problems during the extraction process, guaranteeing that processing would continue even in the event of

corrupted or inconsistent file formats. After the extraction process was finished, a summary report that included information about the total number of movies that were successfully processed was produced. This report allowed for effective tracking and validation of the extraction process. Our project's video-to-audio extraction mechanism was an essential preprocessing step that made it easier to collect the audio data required for the multimodal analysis of human emotions that followed.

## C) DATA PRE-PROCESS:

In our study, we recognized the significance of delving deeper into textual data to enhance our comprehension and modeling skills, especially within the realm of human emotion analysis. To achieve this, we employed an advanced text embedding method, which played a pivotal role in our data preprocessing pipeline. Text embedding served to furnish semantically meaningful representations for the raw textual data in our dataset, facilitating more nuanced analysis and interpretation of human emotions conveyed through the text. Our approach centered on the utilization of cutting-edge methods, with a particular emphasis on the Sentence-BERT (SBERT) model for text embedding. SBERT stands out as an exceptional model for generating high-quality semantic embeddings for textual data. Our aim was to transform textual information into dense vector representations using SBERT, with each vector encapsulating the semantic essence of the corresponding text. Additionally, we undertook an audio feature extraction process to extract pertinent features from audio data. This step proved vital in capturing relevant attributes of the audio signals, thereby enabling a deeper analysis and interpretation of emotional cues embedded within the audio recordings. The audio features extracted encompassed a varied array of descriptors, each capturing specific aspects of the audio signals. These descriptors included: Zero-crossing rate (ZCR): This metric measures the rate at which the audio waveform changes its sign, often indicating the presence of high-frequency components or percussive sounds. Chroma-based features (chroma_stft): These features represent the energy distribution of different pitch classes in the audio signal, facilitating the analysis of musical content and tonal characteristics. Mel-frequency cepstral coefficients (MFCC): These features portray the spectral characteristics of the audio signal, particularly sensitive to human auditory perception. They are commonly employed in speech and audio processing tasks. Root mean square energy (RMS): This metric quantifies the average energy of the audio signal, offering insights into its overall amplitude and loudness. Mel-scale spectrogram (mel): This visualization depicts the frequency content of the audio signal over time, with a focus on the human auditory system's perception of frequency. The emotion that is taken for training are given below: Surprise Disgust sadness Joy Fear Anger Neutral

## D) MODEL BUILDING:

We developed various combinations of modal architectures for both text and audio data. For text data, we employed a Bi-directional Long Short-Term Memory (Bi-LSTM) model with an attention layer. This architecture enables the model to learn patterns and features within the text data effectively. For audio data, we explored two modalities. One modality involved a combination of a Convolutional layer followed by a Bi-directional Long Short-Term Memory (Bi-LSTM) model with an attention layer. The Convolutional layer aids in capturing spatial patterns within the audio features before feeding them into the Bi-LSTM. The other modality excluded the Convolutional layer and solely utilized a Bi-LSTM with an attention layer to learn patterns directly from the audio features.

Both modalities were designed to effectively capture patterns within the audio data.

### E) MODEL FUSION:

Early fusion techniques were employed to merge the outputs from the audio and text models. This fusion process involved using a Flatten layer to flatten the output from each model, followed by concatenation using the Concatenate layer. The merged output was then passed through a final dense layer for emotion prediction. In this architecture, the SoftMax activation function was utilized in the final layer to determine the percentage likelihood of each predicted emotion.

### F) MODEL TRAINING AND EVALUATION:

We utilized the ADAM (Adaptive Moment Estimation) optimizer for training our model, which efficiently adapts learning rates for each parameter during optimization. Additionally, we employed label encoding to represent emotions for training purposes. Among various model architectures, the combination of Bi-directional Long Short-Term Memory (Bi-LSTM) with an attention layer for text data, and a combination of Convolutional layer followed by Bi-directional Long Short-Term Memory (Bi-LSTM) with an attention layer for audio data exhibited superior performance compared to alternative architectures. These architectures effectively captured and learned intricate patterns within both textual and audio data, resulting in enhanced emotion prediction accuracy.

### G) USER INTERFACE CREATION:

When crafting the user interface (UI) for an Emotion Recognition system, simplicity and user-friendliness are paramount to ensure efficient user interaction. Leveraging tools like the open-source Python library Streamlit can streamline the UI design process for deep learning and machine learning models. Streamlit simplifies the creation of intuitive and user-friendly interfaces, enabling seamless interaction with the Emotion Recognition system.

## 4. DATA PREPROCESSING

### 4.1 Text:

Our methodology heavily relies on text embedding to transform raw textual data into a format suitable for deep learning analysis. Leveraging SentenceTransformer, a state-of-the-art library for generating dense vector representations of text, we aim to capture the intricate nuances of meaning present in our dataset of conversation transcripts.

The initial step in our process involves selecting the "all-mpnet-base-v2" model, renowned for its consistent performance across various natural language processing tasks, including text embedding. This choice ensures that our model is built on a solid foundation capable of comprehending and encoding the subtle features of human language effectively. Each textual input is meticulously handled, yielding dense vector embeddings that encapsulate the semantic essence of the corresponding text passage, utilizing SentenceTransformer's encode method. These embeddings serve as rich representations of the source text, encapsulating its contextual, syntactic, and semantic qualities.

This approach aims to condense each conversational utterance into a concise and interpretable numerical form, facilitating subsequent analysis by our deep learning models. To ensure the robustness of our text embedding procedure, comprehensive error handling techniques have been implemented. This ensures the stability of our feature extraction pipeline by gracefully managing unexpected difficulties or anomalies

encountered during text encoding, preserving the integrity and reliability of the process. Furthermore, the seamless integration of our text embedding procedure with the Pandas DataFrame framework streamlines the workflow for incorporating generated embeddings into our data analysis. This integration enables effortless combination of textual features with other data modalities, facilitating multimodal analysis and enhancing the depth of insights derived from our dataset.

### 4.2 Audio:

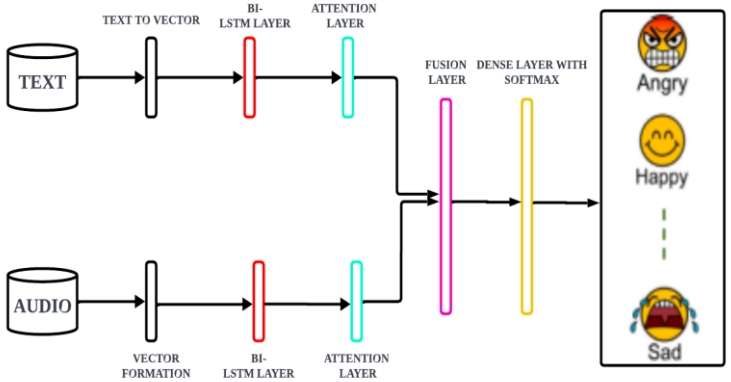
In our project's audio processing pipeline, we employ a two-step approach to extract and process audio data from video clips. First, we utilize the MoviePy library to extract audio streams from MP4 video files, converting them into WAV format for further analysis. This extraction process ensures that we can effectively isolate and manipulate the audio content embedded within each video file. Despite potential challenges such as file format inconsistencies or corrupted files, our robust exception handling mechanisms ensure the continuity of processing, maintaining data integrity throughout the extraction procedure.

After the audio data is retrieved, we use the Librosa library to extract features, which are crucial aspects of the audio signals that are pertinent to the identification of emotions. Zero-crossing rate (zcr), chroma-based features (chroma_stft), root mean square energy (rms), Mel-frequency cepstral coefficients (mfcc) and Mel-scaled spectrogram (mel) are just a few of the descriptors that are included in our feature extraction method. These characteristics give us valuable information about the audio signals' acoustic characteristics, allowing us to identify patterns and cues that correspond to various emotional states.

Our goal with this all-encompassing approach to audio processing is to convert unprocessed audio data into meaningful feature representations that capture the emotional content below. We guarantee the stability and efficiency of our audio processing pipeline by utilising both the MoviePy and Librosa libraries, which paves the way for further multimodal research in conjunction with textual data. Our ultimate objective is to create a comprehensive emotion detection system that can reliably identify human emotions from a variety of modalities. This will improve human-computer interaction and enable applications in a range of fields, including mental health diagnosis and customer service.

## 5. MODEL ARCHITECTURE

The basic overview of our model is it needs to get input in two formats one is text and audio data which is already preprocessed in the process already mentioned above. The text features will run through a pipeline consisting of bi-directional Long-Short term memory and attention layers. For audio features it run through bi-directional long short term memory and attention layers and other modal has Convolutional layer present for better performance in audio parts this output is fused to make a combined feature vector that is passed to a dense layer with SoftMax activation function in the end to predict the similarity of the emotions




In the above image, there are two types of data which are involved in the process. Text data and the audio data are the data, in which the text data is converted to vector representations.

Converting text data into vector representations serves to make textual information amenable to numerical processing, a requirement for artificial learning algorithms. These vectors encapsulate semantic meaning by encoding relationships between words and reducing dimensionality. Consequently, this transformation enables algorithms to efficiently analyze text, extract pertinent features, and undertake various tasks like sentiment analysis, text classification, and natural language understanding. And then it is sent to the BiLSTM layer in which the Bidirectional LSTM (BiLSTM) layer processes input sequences in both backward and forward directions simultaneously. This unique architecture enables the model to capture contextual information from both past and future timesteps. By doing so, it enhances the model's power to understand intricate temporal dependencies and long-range relationships present in sequential data, such as text or time series. By passing through the BiLSTM layer, it reaches the attention layer, where this is crucial in sequence modelling tasks because it allows the model to dynamically focus on relevant parts of the input sequence. During training, the attention mechanism learns to assign weights to different parts of the input sequence, highlighting the most important elements. During inference, these weights guide the model to selectively attend to specific information, enabling more effective processing of long sequences and improving the model's overall performance in tasks such as translation and summarization.

Simultaneously, the audio data is pre-processed, and the below process will happen parallelly. Audio data undergoes conversion into vector representations to facilitate efficient processing and analysis by machine learning algorithms. By representing audio signals as vectors, crucial features like frequency components and temporal characteristics are captured, enabling algorithms to discern significant patterns necessary for tasks such as speech recognition, emotion detection, and audio classification. This transformation streamlines numerical computation processes and ultimately improves algorithm performance. And further the attention layers are pivotal for pinpointing pertinent segments within the input. Since audio signals encompass varied information, specific segments hold more significance for tasks like speech recognition or emotion detection. The dynamic focus on these segments aids in extracting essential features, thereby enhancing accuracy and overall performance in audio-centric applications. The fusion layer integrates processed data from diverse modalities like text, audio, and images, enriching the model's feature representation for more robust predictions. By amalgamating complementary information, it fosters a comprehensive understanding of multimodal inputs, notably enhancing performance in tasks such as emotion recognition. This holistic approach enables the model to capture nuanced relationships across different data types, thereby improving overall accuracy and effectiveness.

The dense layer with a SoftMax activation function serves as the final classification step in the project, computing the probability distribution over different emotion classes based on the fused representation from the fusion layer. This layer enables the model to make predictions regarding the most likely emotion conveyed by the input data, providing a probabilistic assessment of each emotion category and facilitating decision-making in emotion recognition tasks. After this layer the output will be predicted based on prescribed classes of data.

### 5.1 INITIAL MODEL

#### Text Pathway

In our proposed multimodal architecture, text and audio inputs are combined to achieve accurate emotion prediction. The architecture consists of parallel pathways for text and audio analysis, followed by fusion and classification layers.

The text pathway begins with a Bidirectional Long Short-Term Memory (BiLSTM) layer, a powerful variant of recurrent neural networks (RNNs) adept at identifying long-range dependencies in sequential input. The BiLSTM layer comprises two components:

one processes the input sequence forward, while the other processes it backward. This bidirectional processing allows the model to capture contextual information effectively by considering both past and future context. Consequently, the BiLSTM layer excels at understanding the sequential structure of textual input and identifying significant components that accurately represent the nuanced meanings of human speech.

Subsequent to the BiLSTM layer, an Attention Mechanism is employed to emphasize important segments of the text. Attention mechanisms enable the model to focus on relevant information while disregarding noise or irrelevant content. This selective attention mechanism proves particularly beneficial in scenarios where certain segments of the input sequence hold more significance for the task at hand. By assigning higher weights to informative segments, the attention mechanism enhances the model's capability to extract relevant features from the text data. To mitigate overfitting and stabilize the training process, regularization techniques such as Dropout and Batch Normalization layers are integrated into the text pathway. Dropout randomly deactivates a fraction of neurons during training, reducing the model's reliance on specific features and preventing co-adaptation. Batch Normalization normalizes the activations of each layer, rendering the model more resilient to variations in input distributions and accelerating the training process. These techniques ensure that the text pathway remains robust to noise and variations in the input data, thereby enhancing the overall performance of the model.

**Audio pathway**

A BiLSTM layer that is designed to detect temporal patterns in the audio signals starts the audio pipeline concurrently. The BiLSTM layer in the audio path works in tandem with its counterpart in the text pathway, processing input in both directions to extract relavent features from the audio data. The BiLSTM layer successfully captures the dynamics and temporal dependencies present in speech signals by taking into account both past and future context.

The audio pathway applies an Attention Mechanism to the BiLSTM layer's output, just like the text pathway does. By improving the model's capacity to concentrate on key audio segments, the attention mechanism helps it to extract discriminative features that aid in precise emotion prediction. The attention mechanism makes sure that the model pays more attention to salient acoustic cues while reducing the impact of irrelevant or noisy signals by dynamically assessing the value of various audio segments.

The audio pathway incorporates Dropout and Batch Normalization layers to regulate training process and enhance the model's generalization skills. Regularization strategies help reduce the likelihood of overfitting by stabilizing the learning process and adding unpredictable nature to the training process. The audio pathway may acquire stable representations of emotion-related variables thanks to the assistance of Dropout and Batch Normalization layers, which also prevent the model from learning incorrect relationships or noise in the training data. This improves the predictive accuracy of the model.

**Fusion and Classification:**

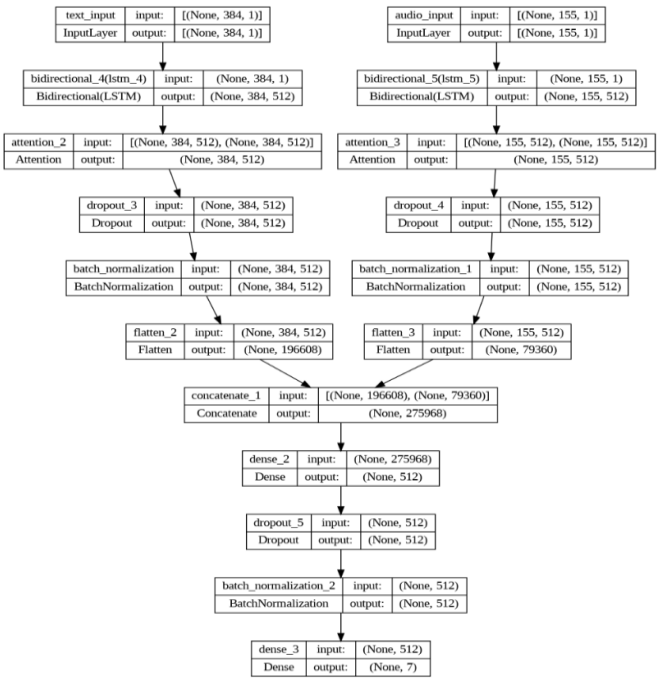
The outputs from the attention layers are concatenated to create a single feature representation after the text and audio routes have processed their respective inputs. A Fully Connected Layer receives this concatenated feature vector as input, which contains the integrated data from both modalities. In order to identify complex patterns and relationships that are suggestive of the underlying emotions expressed in the conversation, the Fully Connected Layer further processes the concatenated data. Lastly, the probabilities for each of the seven emotion classes are predicted by an Output Layer using a SoftMax activation function. The model's predictions can be understood probabilistically thanks to the SoftMax function, which normalizes the output probabilities so that they add up to one. The algorithm can determine the most likely emotion represented in the input discussion and quantify its confidence in each prediction by giving each emotion category a probability. Through this

classification stage, the model is able to classify the input conversation into the relevant emotion class, which helps with downstream applications like affective computing, sentiment analysis, and human-computer interaction.

Model: "model_1"

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| text_input (InputLayer) | [(None, 384, 1)] | 0 | [] |
| audio_input (InputLayer) | [(None, 155, 1)] | 0 | [] |
| bidirectional_4 (Bidirectional) | (None, 384, 512) | 528384 | ['text_input[0][0]'] |
| bidirectional_5 (Bidirectional) | (None, 155, 512) | 528384 | ['audio_input[0][0]'] |
| attention_2 (Attention) | (None, 384, 512) | 0 | ['bidirectional_4[0][0]', 'bidirectional_4[0][0]'] |
| attention_3 (Attention) | (None, 155, 512) | 0 | ['bidirectional_5[0][0]', 'bidirectional_5[0][0]'] |
| dropout_3 (Dropout) | (None, 384, 512) | 0 | ['attention_2[0][0]'] |
| dropout_4 (Dropout) | (None, 155, 512) | 0 | ['attention_3[0][0]'] |
| batch_normalization (BatchNormalization) | (None, 384, 512) | 2048 | ['dropout_3[0][0]'] |
| batch_normalization_1 (BatchNormalization) | (None, 155, 512) | 2048 | ['dropout_4[0][0]'] |
| flatten_2 (Flatten) | (None, 196608) | 0 | ['batch_normalization[0][0]'] |
| flatten_3 (Flatten) | (None, 79360) | 0 | ['batch_normalization_1[0][0]'] |
| concatenate_1 (Concatenate) | (None, 275968) | 0 | ['flatten_2[0][0]', 'flatten_3[0][0]'] |
| dense_2 (Dense) | (None, 512) | 141296128 | ['concatenate_1[0][0]'] |
| dropout_5 (Dropout) | (None, 512) | 0 | ['dense_2[0][0]'] |
| batch_normalization_2 (BatchNormalization) | (None, 512) | 2048 | ['dropout_5[0][0]'] |
| dense_3 (Dense) | (None, 7) | 3591 | ['batch_normalization_2[0][0]'] |

The plot diagram is given below for model:




## 5.2 **UPDATED MODEL:**

The text part for this modal is the same as the above model. We have added the Convolution layer in the audio part to capture the features more accurately.

Layers of Text Processing:

Long Short-Term Memory (LSTM) in bidirectional mode: Two layered BiLSTM layers extract contextual information in both directions from textual inputs. Sequences are processed by the first BiLSTM layer, which has 128 units, and the second BiLSTM layer, which has 64 units. The model is able to effectively learn complex patterns in text input because of its hierarchical structure. Attention Mechanism: By emphasizing noteworthy text passages, an attention layer helps the model concentrate on key data. By dynamically weighting input features according to importance, it enables the model to separate out noise and emphasize meaningful textual clues. Dropout and Flatten Layers: While Flatten layers flatten the output tensors into a single dimension for additional processing, Dropout layers prevent overfitting by randomly removing a portion of units during training.

Layers of Audio Processing:

Convolutional Layers: In our multimodal architecture, Convolutional Neural Networks (CNNs) are essential for extracting discriminative characteristics from audio inputs. We use two Conv1D layers with 64 and 128 filters, respectively, in this situation. Local patterns and spectral characteristics contained in the data are successfully captured by these layers, which use convolution operations throughout the temporal dimension of the audio signals. Each filter in the Conv1D layers acts as a pattern detector, convolving over small segments of the input audio signals to detect local features such as pitch variations, timbre, and transient sounds. By convolving with learned filter weights, these layers effectively extract hierarchical representations of audio features, ranging from low-level spectral components to higher-level abstract representationsMaxPooling1D layers are

employed to downsample the output feature maps. MaxPooling reduces the spatial dimensionality of the feature maps while retaining essential information, thus improving computational efficiency and reducing the model's susceptibility to overfitting.

Bidirectional LSTM: Audio sequences are processed bidirectionally by two stacked BiLSTM layers, which also extract acoustic patterns and temporal dependencies from the audio inputs. Hierarchical feature extraction is made easier by the second BiLSTM layer, which has 32 units instead of the first 64.

Attention Mechanism, Dropout, and Flatten Layers: The audio pathway uses Dropout and Flatten layers for regularization and feature reshaping, as well as an Attention layer to highlight significant audio portions, just like the text pathway.

Fusion and Classification:

Concatenation: To create a single feature representation that captures both the textual and acoustic aspects of conversations, the outputs from the text and audio routes are combined. Fully Connected Layers: The concatenated features are processed by a Dense layer with 512 units and ReLU activation, which captures intricate patterns suggestive of underlying emotions. Output Layer: To enable the model to classify talks into the relevant emotion categories, the last Dense layer, which has a SoftMax activation function, predicts probabilities for each emotion class.

This updated multimodal architecture capitalizes on the complementary strengths of textual and audio modalities, enabling accurate and robust emotion prediction in diverse conversational contexts.

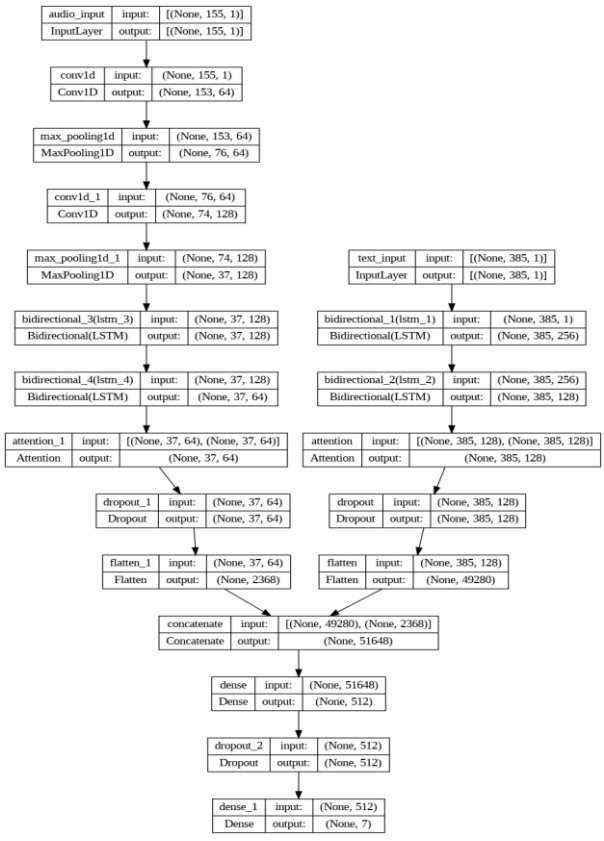
Summary of the model:

Model: "model"

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| audio_input (InputLayer) | [(None, 155, 1)] | 0 | [] |
| conv1d (Conv1D) | (None, 153, 64) | 256 | ['audio_input[0][0]'] |
| max_pooling1d (MaxPooling1D) | (None, 76, 64) | 0 | ['conv1d[0][0]'] |
| conv1d_1 (Conv1D) | (None, 74, 128) | 24704 | ['max_pooling1d[0][0]'] |
| text_input (InputLayer) | [(None, 385, 1)] | 0 | [] |
| max_pooling1d_1 (MaxPooling1D) | (None, 37, 128) | 0 | ['conv1d_1[0][0]'] |
| bidirectional (Bidirectional) | (None, 385, 256) | 133120 | ['text_input[0][0]'] |
| bidirectional_2 (Bidirectional) | (None, 37, 128) | 98816 | ['max_pooling1d_1[0][0]'] |
| bidirectional_1 (Bidirectional) | (None, 385, 128) | 164352 | ['bidirectional[0][0]'] |
| bidirectional_3 (Bidirectional) | (None, 37, 64) | 41216 | ['bidirectional_2[0][0]'] |
| attention (Attention) | (None, 385, 128) | 0 | ['bidirectional_1[0][0]', 'bidirectional_1[0][0]'] |
| attention_1 (Attention) | (None, 37, 64) | 0 | ['bidirectional_3[0][0]', 'bidirectional_3[0][0]'] |
| dropout (Dropout) | (None, 385, 128) | 0 | ['attention[0][0]'] |
| dropout_1 (Dropout) | (None, 37, 64) | 0 | ['attention_1[0][0]'] |
| flatten (Flatten) | (None, 49280) | 0 | ['dropout[0][0]'] |
| flatten_1 (Flatten) | (None, 2368) | 0 | ['dropout_1[0][0]'] |
| concatenate (Concatenate) | (None, 51648) | 0 | ['flatten[0][0]', 'flatten_1[0][0]'] |
| dense (Dense) | (None, 512) | 26444288 | ['concatenate[0][0]'] |
| dropout_2 (Dropout) | (None, 512) | 0 | ['dense[0][0]'] |
| dense_1 (Dense) | (None, 7) | 3591 | ['dropout_2[0][0]'] |

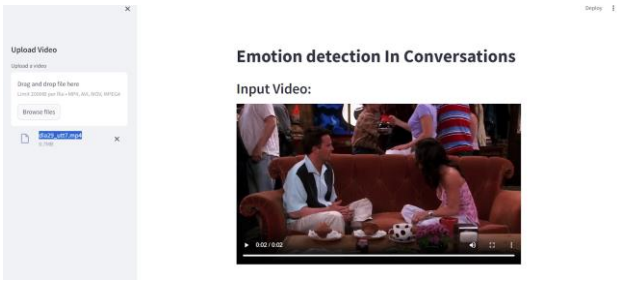
Total params: 26910343 (102.65 MB)
Trainable params: 26910343 (102.65 MB)
Non-trainable params: 0 (0.00 Byte)

The plot diagram for the model:

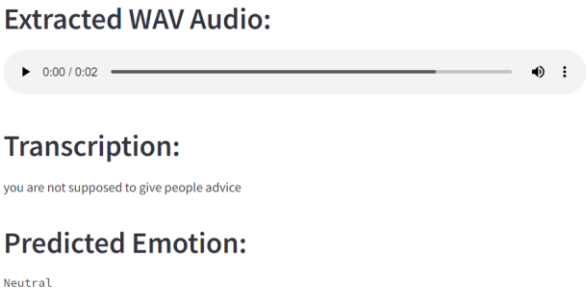


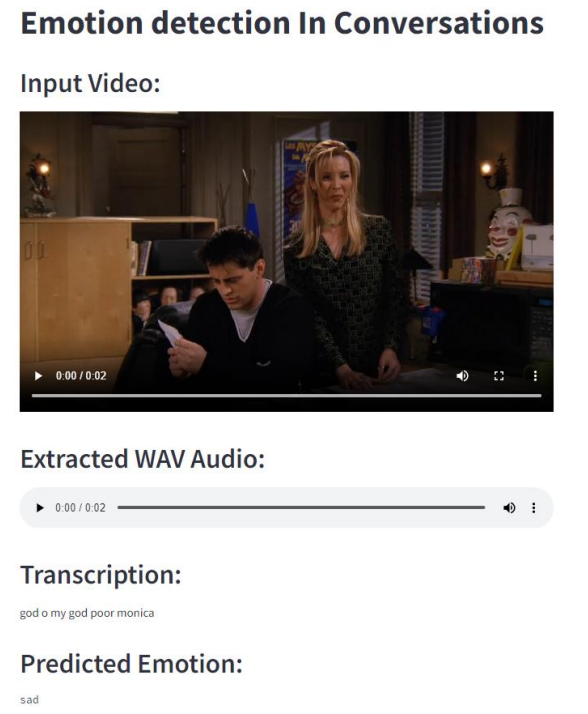
## 6. RESULT

We have used to Streamlit library to build the User interface for getting the Input video from the user and made prediction on the given input by the model we have developed. The User Interface is shown below:




We will first give the input Conversation Video as an input and the ui will load the input and display it.




The initial model that is trained using the balanced MELD Dataset got a very low training accuracy of 25 % and when we trained with full MELD dataset we got a training Accuracy of 47 %. The updated model got a training accuracy of 55% and performs well compared to the initial model.

### 6.1 Significance, Strengths and Limitations:

The proposed multimodal emotion recognition system stands at the forefront of technological innovation, offering unprecedented potential to revolutionise various sectors including customer service, user engagement, and healthcare. By harnessing the power of multiple data modalities such as text, audio, and images, this system transcends traditional boundaries to provide a holistic understanding of emotional expressions. This integration enables a more nuanced interpretation of human emotions, allowing for more personalised and empathetic interactions with users.

One of the system's key strengths lies in its ability to seamlessly integrate and process diverse types of data. Textual inputs offer rich contextual information, while audio signals capture intonation and vocal cues, and images convey facial expressions and body language. By combining these modalities, the system can derive deeper insights into the emotional states of individuals, leading to more effective communication and decision-making processes.

However, despite its promise, the system faces several challenges that must be addressed to realise its full potential. Ensuring the accuracy of emotion recognition algorithms remains a critical concern, as misinterpretations could lead to misunderstandings or inappropriate responses. Additionally, maintaining user privacy is paramount, especially when dealing with sensitive emotional data. Robust privacy

measures and data encryption techniques must be implemented to safeguard user information.

Furthermore, the system must navigate the complexities of cultural nuances and individual differences in emotional expression. Emotions are deeply influenced by cultural norms, societal expectations, and personal experiences, making it essential for the system to be adaptable and culturally sensitive. This requires ongoing research and development efforts to ensure that the system's algorithms are capable of recognizing and accommodating diverse expressions of emotion.

Despite these challenges, the potential benefits of the proposed system are immense. By fostering more empathetic human-machine interactions, the system has the power to enhance user experiences, improve healthcare outcomes, and drive innovation across various industries. Through continuous refinement and a commitment to ethical standards, this system can pave the way for a future where technology truly understands and responds to human emotions in meaningful ways.

## 7. CONCLUSION:

The proposed model architecture with fusion technique works well with bi-lstm and attention layer on text and conv layers + bi-lstm and attention on audio side. We though we only got 55% training accuracy the model is trained on MELD dataset which has data from clipped from real word TV shows which represent the real-world scenario of Human speaking. By training and increasing the model with real world scenario data will helps model to prediction human emotions in real time conversations and can be best useful for Human Computer Interactions in future.

## 8. REFERENCE


- T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," Natural Language Processing Journal, 2023.
- J. Mutinda, W. Mwangi, and G. Okeyo, "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network," Applied Science Article, vol. 2023, Jan. 2023.
- In IEEE format, the reference would appear as follows: M. Kalpana Chowdary, T. N. Nguyen, and D. Jude Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," Neural Computing and Applications, vol. 35, no. 2023, pp. [page numbers], Apr. 2021.
- M. Jafari, A. Shoeibi, and M. Khodatars, "Emotion recognition in EEG signals using deep learning methods: A review," Elsevier Journals, 2023.
- Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning," 2023.
- Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," Elsevier, 2023. [Online].
- HONG-WEI. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," Advanced Digital Sciences Center (ADSC), University of Illinois

at Urbana-Champaign, Singapore, [Emails: hongwei.ng, vietdung.n, bbonik, stefan.winkler]@adsc.com.sg.

- 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, 2023.

- K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," Artificial Intelligence, Data Engineering and Applications Laboratory, Faculty of Science of Bizerte, University of Carthage, Zarzouna 7021, Tunisia.