Emotion Recognition in Speech by Multimodal Analysis of Audio and Text

Siddhant Bikram Shah
Computer Science and Engineering
Delhi Technological University
New Delhi, India
siddhantshah3000@gmail.com

Shubham Garg
Computer Science and Engineering
Amity University
Noida Uttar Pradesh, India
shubgarg17@gmail.com

Aikaterini Bourazeri
Computer Science and Electronic Engineering
University of Essex
Colchester, United Kingdom
a.bourazeri@essex.ac.uk

2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-6263-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/Confluence56041.2023.10048872

***Abstract*—Emotion recognition remains a very challenging task in research because of its sensitive and multifaceted nature. Recently, emotion recognition has garnered a lot of attention owing to its significance in psychology, human-computer interaction, and healthcare, where people's facial expressions, voice qualities, and spoken words are used to better understand it. While emotion recognition holds the power to facilitate various health problems, the main challenge emotion recognition systems face is to accurately identify hidden nuances in expressions and thus, the underlying emotions conveyed by them. The true emotions of a person may remain concealed or not properly identified when only one mode of input is analyzed, therefore, multimodal streams of inputs are used to provide a more holistic view of a person's emotions. In this paper, a novel framework that fuses the results of two uni-modal methods of emotion recognition, audio, and text, to develop a robust and versatile emotion recognition system is proposed. The results show that signal processing and language processing can be utilized to reliably detect emotion from audio and text, with an accuracy of 96% and 94.1% respectively. Further, the approach presented in this paper can be used as a depression detection and monitoring tool to further enable mental healthcare professionals accurately detect symptoms of depression.**

***Index Terms*—Affective Computing, BERT, Deep Learning, Emotion Recognition, Multimodality**

I. Introduction

Humans have the innate ability to recognize the emotions conveyed by other humans. Emotion recognition drives interaction between people and is crucial in social environments for mutual understanding, gauging affability, and avoiding conflict. Humans are very revealing by nature, as they disclose emotions in many ways [1]. For example, even something as seemingly uninvolved as eyebrows can convey a multitude of emotions [2]. While being taken for granted due to its intrinsic nature, the potential of emotion recognition in a wide variety of applications cannot be understated.

Emotion recognition is a challenge being undertaken by a plethora of multidisciplinary researchers owing to its relevance in psychology, human-computer interaction, affective computing, and healthcare [3]. Recent developments in computer vision, natural language processing, and signal processing have brought new life to these efforts, which can be ultimately attributed to the rise of deep learning. Many researchers exploring emotion recognition have been trying to answer the same question: can machines achieve or even surpass human-level emotion recognition?

Human speech is a complex bundle of data comprising multiple modalities, each having its own features, making it a very widely used data medium in the deep learning community, especially in emotion recognition [4]. From speech data, the words spoken by a person and the innate features of their voice can be extracted. These modalities prove to be descriptive and reliable indicators of emotions when processed by deep learning techniques, especially when used in tandem.

The words spoken by a person can be stored as text. Thereafter, analyzing the emotion in the text becomes a language-processing problem. This text, however, is not very realistic when it comes to the recognition of a person's emotion in real-life situations [5]. For instance, when a person uses sarcasm, they may use words that convey a completely different picture when another context is not considered. Therefore, while text data is a powerful medium for predicting emotions, it suffers from many problems that leave much to be desired for accurate emotion recognition. To move towards more realistic and accurate emotion recognition, relevant additional context must be considered by intelligent systems.

The audio signals within speech data can be analyzed to recognize the speaker's emotions. While spoken words are the most direct indicators of emotions, there is more to the picture. This can be clearly seen in the case of affective animals like dogs, who consider the underlying tone of human speech and put less weight on the words spoken [6]. This is a testament to the natural instincts of creatures when it comes to sound. However, acoustic data does not come with its shortcomings, as it has a high probability of noisy data and is easily skewed by the innate voice qualities of people.

The inadequacies of the aforementioned modalities make it apparent that one single modality is not enough to represent this complex real-world problem. The combination of multiple modalities, however, may allow them to account for each other's limitations. Furthermore, it can be used to obtain a better view of the bigger picture, which may lead to a better solution. In the context of emotion recognition, multimodality is a powerful tool that allows intelligent systems to recognize the emotions of a person more accurately.

In this study, a multimodal approach for emotion recognition

is proposed that predicts emotion in speech data through unimodal analysis of text and audio modalities and fusion of the results obtained. A detailed review of the recent developments in emotion recognition is presented in Section II. Section III describes our methodology and all its constituents in detail. In Section IV, the results of our experiments are presented. A discussion along with the future directions of our work is presented in Section V. Finally, Section VI concludes our study.

II. Related Work

A. Emotion Recognition

Until recently, most of the work regarding emotion recognition was based on unimodal data. This section presents a review of various multimodal approaches for emotion recognition.

A self-assessment tool was proposed by Prabhu et al. [7] that uses audio, video, and text input for emotion recognition and ultimately, depression screening. The tool uses a one-of-three scheme, and therefore if one modality indicates that the person is depressed, the person is classified as depressed.

Wu et al. [8] proposed a framework to detect depression based on recognizing emotions conveyed through audio and text inputs. This work explored the association between depression and emotion recognition, and how they influence each other. However, they did not subcategorize the severity of depression of the individual.

A novel framework, the cLSTM-MMA, was proposed by Pan et al. [9] to aid speech emotion recognition using video and text. Multimodal information from visual and textual cues helps to alleviate the ambiguity in emotion recognition via speech.

The work by Luna-Jimenez et al. [10] showed that audio and video carry enough separate but relevant details that when combined, the accuracy of emotion recognition systems is greatly improved. This could also be applicable to the text modality, however, this study did not explore it.

Peng et al. [11] proposed MSCNN-SPU-ATT for emotion recognition by extracting, attending, and fusing audio and text features. This simple system forwent the implementation of LSTM networks for computationally simpler CNNs and still achieved competitive results.

Chamishka et al. [12] proposed a novel method for extracting Bag-of-Audio-Words (BoAW) based features for six basic emotions from the IEMOCAP dataset. They proposed an RNN (Recurrent neural network) based architecture for real-time conversation emotion recognition. This architecture was able to outperform several previous state-of-the-art models and achieved a weighted accuracy of 60.87% and an unweighted accuracy of 60.97%.

Sowmya et al. [13] experimented with several machine learning models like SVM, RF and Decision Tree and an MLP (Multi-Layer-Perceptron) for distinguishing between 4 basic emotion types: happy, neutral, sad, and angry. They were able to achieve an accuracy of 85% on their dataset using these traditional approaches. However, the advantages of this model could not be validated using only these 4 classes proposed by the authors. Better results could be obtained using newer methods such as the BERT transformer.

Krishna et al. [14] experimented with various traditional machine learning models over 8 classes: happy, sad, angry, fearful, disgusted, calm, neutral, and surprised. They achieved an accuracy of 86.5% which is slightly better than the results obtained by the previous study.

Finally, Susithra et al. [15] proposed an architecture to recognize the speaker's emotions and gender. For emotion detection, the model was trained to recognize 4 basic emotion types. A simple Feed-Forward Neural Network (FNN) was used for the gender recognition block and a Convolutional Neural Network (CNN) with a good dataset was used for the emotion detection block. The proposed model obtained an accuracy of 91.46% on the gender detection part and 86% on the emotion recognition block.

B. Our Contribution

While a lot of research has been conducted on emotion recognition using audio-visual modalities, the multimodal analysis of audio and text towards the same goal remains relatively unexplored, despite the combination of these two modalities being more seamless. Therefore, this work explores a novel multimodal fusion framework that aims to account for some of the shortcomings of prior work on this topic by using a combination of custom and state-of-the-art detection architectures, using more emotion categories and equal consideration of both modalities.

III. Methodology

A. Text

For the classification of textual data, feature vectors are extracted from a sequence of words and fed into a classifier. This framework is to be used on common phrases often encountered in regular conversations between humans. Therefore, to train our model, an NLP dataset [16] containing sentence-emotion pairs was used. Training the model on such a dataset enabled us to predict the emotions conveyed by a person through their words. This dataset has six emotion categories: joy, sadness, anger, fear, love, and surprise.

For data augmentation, all the sentences in our dataset were parsed through a translation service, converting them to one of 5 random languages, and then back to English. Owing to these new data samples, the number of samples in our dataset increased from 16,000 to 96,000, a 500% increase. In addition, the samples that were translated back to English contained an assortment of synonyms in the place of the original words, enabling our model to associate a wide variety of words with their corresponding emotions, just by increasing its vocabulary. This also allowed our model to become more robust to possible syntactical and semantical errors in the people's sentences.

The invention of transformers has given new life to language processing research as they possess advanced language understanding capabilities. Bidirectional Encoder Representations (BERT) [17] is a family of transformers that have consistently

achieved state-of-the-art performance in text classification problems since its invention. Models trained on the BERT paradigm possess a general understanding of the language, making fine-tuning much faster. For the text processing modality, the proposed framework uses the pre-trained BERT-base-uncased model from the Huggingface library. Transfer learning was implemented by fine-tuning it in our augmented text dataset. The model takes a sentence as input and recognizes the emotion conveyed by it.

B. Audio

The classification of audio data is mainly done by converting analog data into a digital format such as a mel spectrogram and then feeding the binary representation of that data into a deep learning classifier. This data consists of a multitude of valuable features that the classifier can correlate with emotion. For this study, five audio features from the audio samples were extracted:

- Chroma Short-Time Fourier Transform (Chroma STFT) [18]: Chroma features contain the harmonic and tonal information of an audio signal.
- Mel Frequency Cepstral Coefficient (MFCC) [19]: MFCCs are a set of features within an audio wave that describes the structure of a spectral envelope.
- Mel Spectrogram [20]: A mel spectrogram represents the different frequencies of an audio wave in the mel scale.
- Root Mean Square Value (RMS Value) [21]: The RMS value of a signal indicates the steady state of energy possessed by that signal. It scales with the amplitude of the signal.
- Zero Crossing Rate (ZCR) [22]: The ZCR of a signal is used as an indicator of the smoothness of the audio signal.

To generalize our model by using a wide variety of data, five different speech datasets were combined to train our audio classification model. The combined dataset has seven emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. Table I presents a summary of the datasets used.

TABLE I
COMPARATIVE ANALYSIS OF THE METHODS IN TERMS OF ACCURACY

Dataset	No. of actors	No. of utterances
Crema-D [23]	91	7442
EmoDB[24]	10	535
RAVDESS[25]	24	1440
SAVEE[26]	4	480
TESS[27]	2	2800

In real-world applications, background noise or disturbances accompany audio recordings. To account for realistic acoustic conditions, the noise was added to the original audio sample, and the pitch and speed of the audio recording were further modified. Additionally, time shift was performed on the audio wave. These modifications also accomplished data augmentation, producing additional data samples to train the model. In total, 48,649 samples were used to train the model.

Convolutional Neural Networks (CNN) are a subclass of deep learning architectures that can be applied to a wide variety of data. CNNs are used to recognize the underlying patterns in sequential data, and thus are very useful in computer vision and signal processing. For the classification of audio, our framework uses a 1-D CNN to process the features present in audio data. The model takes in audio features and predicts the emotion conveyed by them.

The proposed audio emotion detection model is 21 layers deep and consists of a total of 6 types of layers namely: Convolution1d layers for feature extraction, batch normalization layers to normalize the outputs before passing to the next layer, MaxPooling1D for downsampling the feature maps, dropout layer for preventing overfitting of the model, flatten layer for flattening the final output and the dense layer for getting the final output. The model is a comparatively lightweight model when compared to other state-of-the-art architectures and has only 1,674,375 trainable parameters. The inference time of the model is also almost real-time, making it perfect for the task. The Adam optimizer with a learning rate of 0.001 was used to train the model and the loss function used was categorical cross-entropy loss. Figure 1 describes the architecture of our audio classification system.

While text and audio are powerful modalities that convey emotions in various ways, they can individually fall short of capturing the entire picture. This is especially true when dealing with emotion recognition, which is a complex and multifaceted real-world problem. One solution to overcome this problem is multimodality: the combination of multiple streams of data. Multimodality provides rich training signals that help to make models more robust.

To overcome the respective limitations of the text and audio modalities, a multimodal emotion recognition framework was developed. The principle behind this idea is that the audio and text modalities would complement each other by providing different data that accounts for each other's shortcomings. Our framework processes audio and text by using separate classification models and then fuses the results obtained from them in a weighted average model to obtain the final emotion class.

The working process of our framework is described below:

- Speech data: Speech data is fed to our framework as input. The speech data is stored in the form of .wav files and uploaded to the AWS cloud for storage in an AWS S3 bucket.
- Preprocessing: For the text classification system, AWS transcribe, a speech-to-text service, is used to process speech data. Transcribe takes the speech data file as input and reproduces the words spoken in the speech file in the form of text sentences. Thereafter, the feature vectors from the text sentence(s) are extracted by the BERT tokenizer.

  The speech file is directly fed to the audio feature extractor that uses the Librosa library to extract various audio features (Chroma STFT, MFCC, Mel Spectrogram, RMS Value, ZCR) from the speech data.

- Processing: The text feature vectors are fed into our BERT classifier that predicts the underlying emotion in the given text. Similarly, a 1-D CNN is used to process the audio features and produces the emotion conveyed by them. The outputs from both of these systems are obtained individually in the form of probability distributions that represent the probability of each emotion.

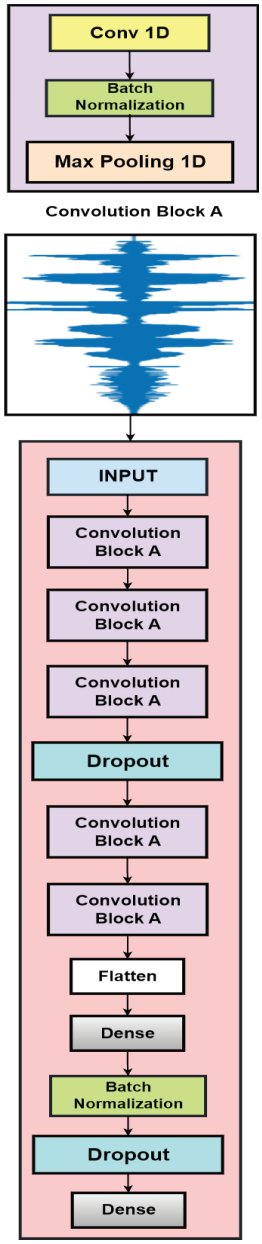

Fig. 1. Architecture of our audio classification CNN.

- Fusion: The outputs obtained from the classification models are fused by using late fusion. A weighted average model aggregates both inputs and produces one final output – the emotion most probably conveyed in the initial speech data.

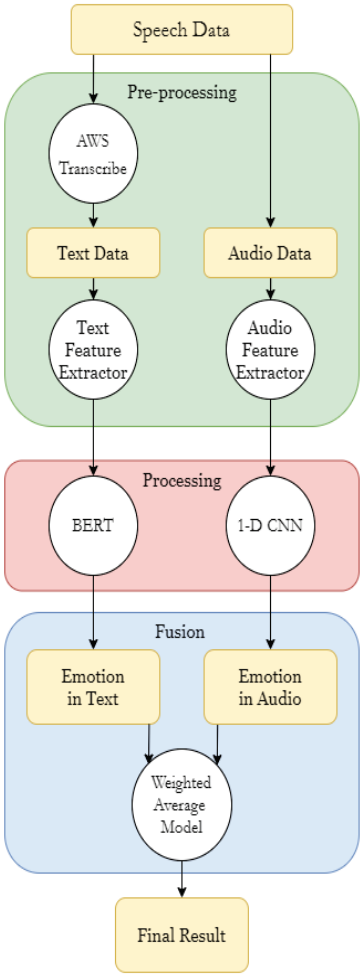

Fig. 2. Flow diagram of our proposed framework.

Figure 2 represents the flow diagram of our proposed framework which combines the text and the audio modalities using a weighted average metric.

IV. Results

A. Text

The text classification system, which uses a BERT transformer, succeeded at classifying emotions with an accuracy of 94.1% and an F-1 score of 0.94. The augmentation and preprocessing of the text dataset improved on the previously obtained accuracy of 91.8%. The model was trained with a learning rate of 0.00002 and a weight decay of 0.01. The batch size for the training was chosen as 8 based on several experiments. The proposed model outperformed most previous state-of-the-art models by a significant margin and employs a fairly novel architecture.

According to our experiments, the pre-trained BERT-base-uncased model performed better than the smaller language models like the Naive Bayes Classifier. BERT transformer is a very powerful and large-weight model, therefore it works

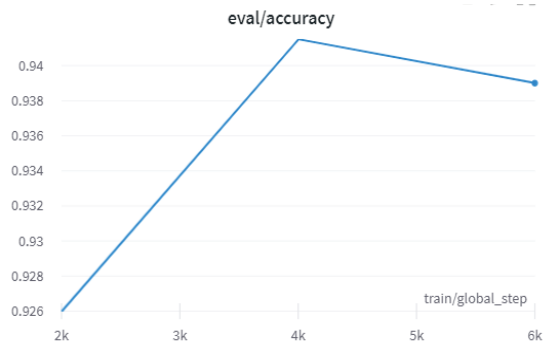

Fig. 3. BERT training curve

great in almost all NLP scenarios but suffers from a huge training time. As it can be inferred from Figure 3, two epochs were enough to make the model converge, owing to the large number of samples in our dataset. Any further epochs caused overfitting, resulting in lowered evaluation accuracy. Figure 4 presents the normalized confusion matrix for the text classification categories.

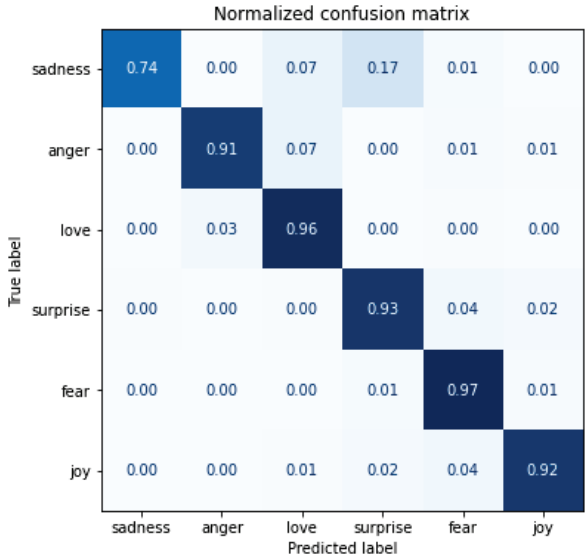

Fig. 4. Normalized confusion matrix of the text modality.

B. Audio

The proposed audio classification model (1D CNN) achieved a validation accuracy of 96% and a validation loss of 0.1095 with an F-1 score of 0.964. Also, the total precision and recall for the model were 0.965 and 0.964 respectively. The model was trained for a total of 250 epochs and reached its top accuracy within the first 180 epochs. As shown by Figure 5, the model did not have any overfitting, and it converged quite smoothly with minimal fluctuations.

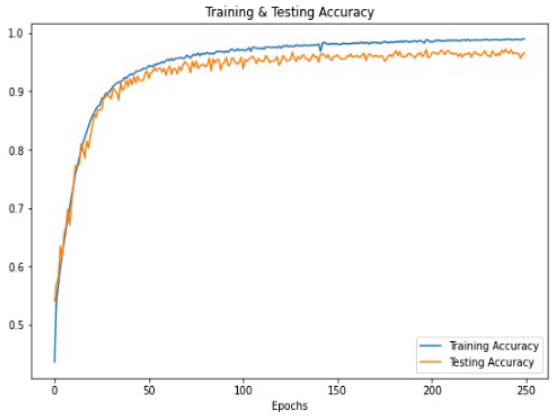

Fig. 5. CNN training curve

TABLE II
CLASSIFICATION REPORT FOR AUDIO (1D CNN)

Class Name	Precision	Recall	F1 Score
Angry	0.96	0.98	0.97
Disgust	0.96	0.95	0.95
Fear	0.94	0.96	0.95
Happy	0.97	0.95	0.96
Neutral	0.98	0.96	0.97
Sad	0.96	0.97	0.96
Surprise	0.99	0.98	0.99
Total	0.965	0.964	0.964

Figure 6 and Table II present the confusion matrix and the classification report respectively for the audio classification categories. It can be inferred from the table that the model performed well over all classes with a maximum F-1 score for surprise (0.99) and a minimum F-1 score for disgust and fear (0.95). The proposed model classified all 7 classes accurately with minimal errors.

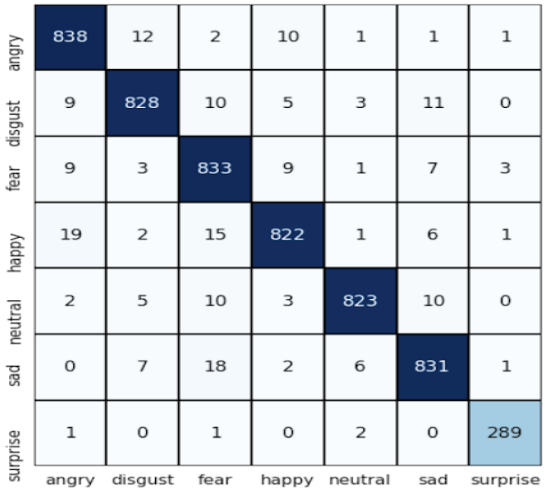

Fig. 6. Confusion matrix of the audio modality.

Moreover, the training time for the 1D CNN model was under 85 seconds for training 41350 audio samples which is minimal when compared to other state-of-the-art architectures. This makes it a reliable, lightweight and low latency model choice for the task of audio emotion classification, and perfect choice to be used with BERT as its text-emotion detection counterpart is a heavyweight model and likely to take up more computation power and processing time.

V. Discussion and Future Work

There is a plethora of emotion recognition datasets available for researchers. The proposed framework is meant to be used on real-life speech dialogues, therefore, it is trained using the best-suited datasets, for the training of the audio classification system as well as the text classification system.

In the text dataset, the number of samples for the emotion 'joy' outnumbers the other emotion categories by a significant margin. To address this problem, multiple datasets can be combined to balance the number of samples in each emotion class. Furthermore, while BERT is very good for text classification tasks, newer text classification models, some from the BERT family, have been engineered to excel at sentiment analysis. Experimentation with these specific models may lead to even better results.

The audio classification part of our framework uses a simple CNN, as in our experiments, heavier models tended to overgeneralize data points and lead to less accuracy. However, due to the large number of data samples used in our experiments, a simple CNN might not be able to effectively capture the scope of the entire data. Therefore, further experimentation might reveal an optimal CNN size that addresses both problems outlined above. In addition, some audio features may increase the complexity of the data but may lead to better results for the audio classification task.

The proposed framework uses more than one modality to achieve realistic emotion recognition, a combination of text and audio modalities. However, introducing more modalities to our framework might produce an even more accurate depiction of a person's true feelings. A vision-based emotion recognition system that operates by using facial movement markers to detect emotions, will make this framework even better.

Persistent pessimistic emotions are considered a symptom of depression. As the next step, this emotion recognition framework can be utilized as an early diagnosis method for depression. Expert advice from mental health professionals, linking the relationship between emotions and depression, will be consulted in order to facilitate this transition. The proposed diagnosis system would be later integrated with a counselling chatbot, in order to achieve our goal of building a private, convenient and non-invasive method for depression detection and therapy.

VI. Conclusion

Emotion recognition is a very delicate yet complex field. While it is useful in many fields such as healthcare, psychology, and human-computer interaction, most existing work on this subject is straightforward and does not consider the multitude of complications and abnormalities that may prevent accurate prediction in real-life scenarios.

This study proposes a speech emotion recognition framework by combining inferences from two different modalities: audio and text. The words spoken and the audio signals within speech data are processed by the framework to obtain the underlying emotion conveyed by them. Various augmentation and preprocessing methods were applied to the datasets in order to train our framework for realistic situations. Our framework uses separate classification systems for predicting emotions in text and audio. The text classification system uses the BERT model, and the audio classification system uses a 1-D CNN model. The text and audio models achieved accuracies of 94.1% and 96%, and F-1 scores of 0.94 and 0.96 respectively. The results obtained from these models are then fed into a weighted average model that combines both of them to produce the most probable emotion conveyed in the given speech data as the result.

Finally, in future work, this emotion recognition system can be translated into a diagnosis system for depression by analyzing the connection between emotions and depression.

References

[1] Schirmer, Annett, and Ralph Adolphs. "Emotion perception from face, voice, and touch: comparisons and convergence." Trends in cognitive sciences 21.3 (2017): 216-228.

[2] Barrett, Lisa Feldman, et al. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements." Psychological science in the public interest 20.1 (2019): 1-68.

[3] Vinola, C., and K. Vimaladevi. "A survey on human emotion recognition approaches, databases and applications." ELCVIA: electronic letters on computer vision and image analysis (2015): 00024-44.

[4] Akçay, Mehmet Berkehan, and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers." Speech Communication 116 (2020): 56-76.

[5] Boukes, Mark, et al. "What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools." Communication Methods and Measures 14.2 (2020): 83-104.

[6] Albuquerque, Natalia, et al. "Dogs recognize dog and human emotions." Biology letters 12.1 (2016): 20150883.

[7] Prabhu, Sahana, et al. "Harnessing emotions for depression detection." Pattern Analysis and Applications 25.3 (2022): 537-547.

[8] Wu, Wen, Mengyue Wu, and Kai Yu. "Climate and Weather: Inspecting Depression Detection via Emotion Recognition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[9] Pan, Zexu, et al. "Multi-modal attention for speech emotion recognition." arXiv preprint arXiv:2009.04107 (2020).

[10] Luna-Jiménez, Cristina, et al. "Multimodal emotion recognition on RAVDESS dataset using transfer learning." Sensors 21.22 (2021): 7665.

[11] Peng, Zixuan, et al. "Efficient speech emotion recognition using multi-scale cnn and attention." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[12] Chamishka, Sadil, et al. "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling." Multimedia Tools and Applications 81.24 (2022): 35173-35194.

[13] Sowmya, G., et al. "Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS Dataset." 2022 International Conference on Electronics and Renewable Systems (ICEARS). IEEE, 2022.

[14] Vamshi, Krishna B., Ajeet Kumar Pandey, and Kumar AP Siva. "Topic model based opinion mining and sentiment analysis." 2018 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2018.

[15] Susithra, N., et al. ”Speech based Emotion Recognition and Gender Identification using FNN and CNN Models.” 2022 3rd International Conference for Emerging Technology (INCET). IEEE, 2022.
[16] Praveen (2019). Emotions dataset for NLP, Version 1. Retrieved August 21, 2022 from https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp.
[17] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).
[18] Müller, Meinard. ”Short-Time Fourier Transform and Chroma Features.” Lab Course, Friedrich-Alexander-Universität Erlangen-Nürnberg (2015).
[19] Likitha, M. S., et al. ”Speech based human emotion recognition using MFCC.” 2017 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE, 2017.
[20] Yenigalla, Promod, et al. ”Speech Emotion Recognition Using Spectrogram Phoneme Embedding.” Interspeech. Vol. 2018. 2018.
[21] Bhat, Aathreya S., et al. ”An efficient classification algorithm for music mood detection in western and hindi music using audio feature extraction.” 2014 fifth international conference on signal and image processing. IEEE, 2014.
[22] Aouani, Hadhami, and Yassine Ben Ayed. ”Speech emotion recognition with deep learning.” Procedia Computer Science 176 (2020): 251-260.
[23] Cao, Houwei, et al. ”Crema-d: Crowd-sourced emotional multimodal actors dataset.” IEEE transactions on affective computing 5.4 (2014): 377-390.
[24] Burkhardt, Felix, et al. ”A database of German emotional speech.” Interspeech. Vol. 5. 2005.
[25] Livingstone, Steven R., and Frank 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 13.5 (2018): e0196391.
[26] Jackson, Philip, and SJUoSG Haq. ”Surrey audio-visual expressed emotion (savee) database.” University of Surrey: Guildford, UK (2014).
[27] K. Dupuis and M. K. Pichora-Fuller, Toronto Emotional Speech Set (TESS). University of Toronto, Psychology Department, 2010