MULTIMODAL EMOTION DETECTION

Submitted By:
Tarun Saxena (002979327)
Kush Suryavanshi (002724729)
Prakriti Pritmani (002724130)

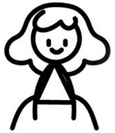
INTRODUCTION

Emotion recognition using audio is a rapidly growing field of research that aims to develop algorithms that can automatically detect and classify emotions in speech.

This is a challenging task due to the complex nature of human emotions and the variation in how they are expressed across individuals and cultures.

Multimodal emotion recognition is the process of detecting and interpreting emotional cues from multiple sources or modalities, such as facial expressions, body language, and speech.

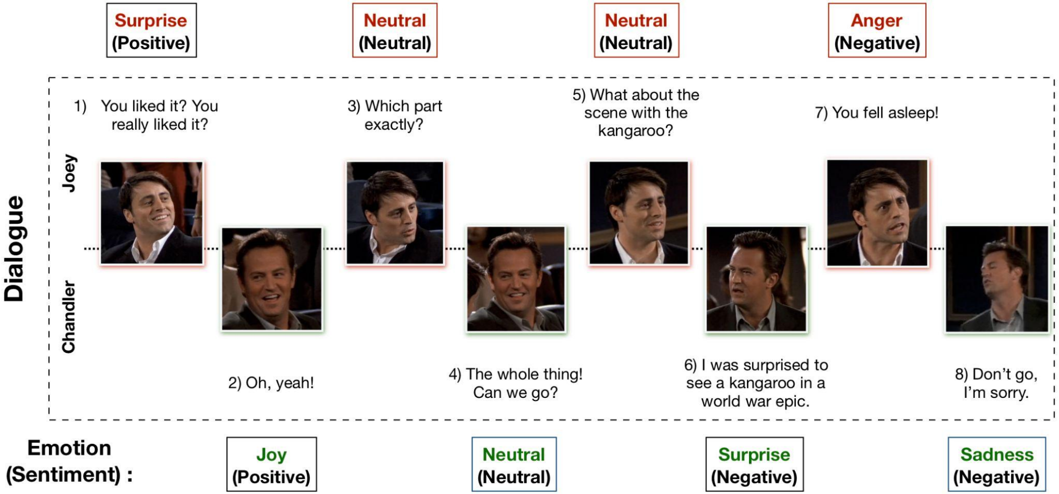
This is an important area of research in the field of affective computing, as it has potential applications in areas such as mental health, education, and social robotics.



DATASET

Multimodal Emotion Lines Dataset (MELD) has been created by enhancing and extending EmotionLines dataset. MELD includes audio and visual media in addition to text and includes the same dialogue instances as EmotionLines. More than 1400 dialogues and 13000 utterances from the Friends TV series are available in MELD. Each utterance in a dialogue has been labeled by any of these seven emotions :

- **Anger**
- **Disgust**
- **Sadness**
- **Joy**
- **Neutral**
- **Surprise**
- **Fear**

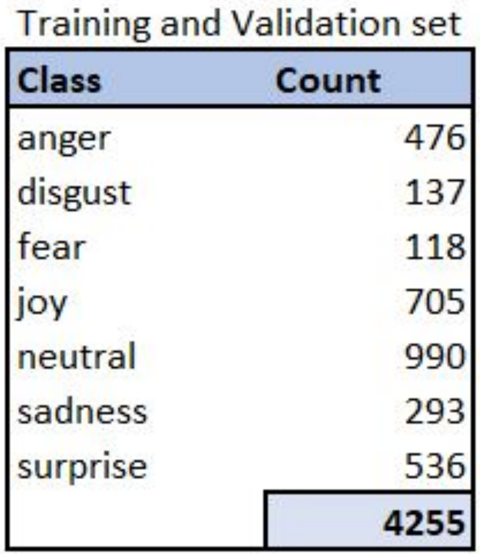


AUDIO ANALYSIS

Audio files need to be preprocessed depending on the architectures to predict the associated emotion.

Two methodologies used to classify emotion and derive insights from:

1. Support Vector Machine (SVM)
2. Time Distributed CNN



Training and Validation set

Class	Count
anger	476
disgust	137
fear	118
joy	705
neutral	990
sadness	293
surprise	536
	4255

LITERATURE REVIEW

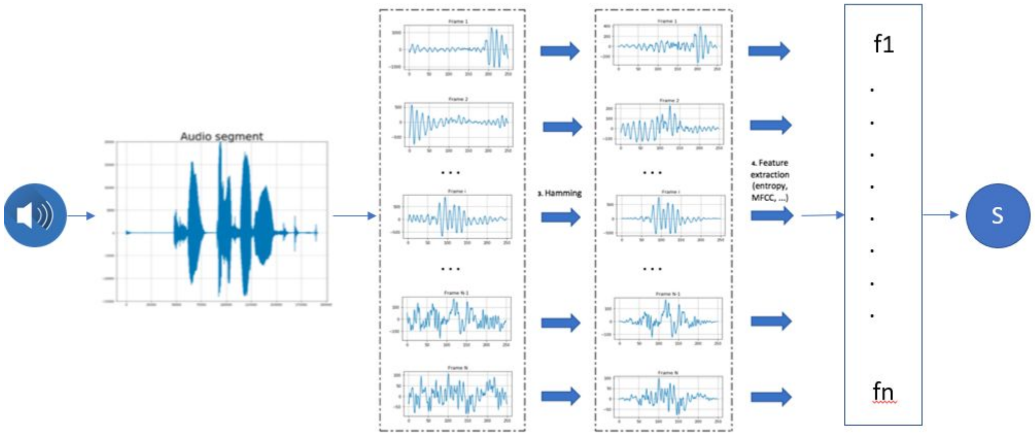
- Dario B et al propose a real time CNN model that is trained from raw audio on a small dataset of TED talks speech data, manually annotated into three emotion classes. They have compared their methodology to other models such as SVM and have claimed to be able to reach better results.

- Nivedita P et al. propose a compact representation of audio using conventional autoencoders for dimensionality reduction as input to multiple different architectures such as decision trees and pre-trained cnns to allow emotion detection .

- Wei G et al. propose a paper that uses two classification methods, the hidden Markov model (HMM) and the support vector machine (SVM), to classify five emotional states: anger, happiness, sadness, surprise and a neutral state

AUDIO ANALYSIS

SVM

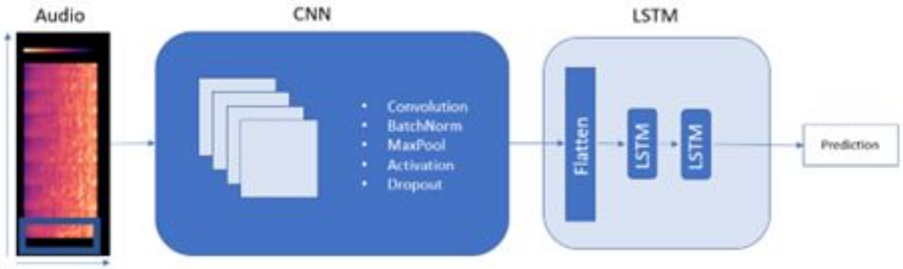


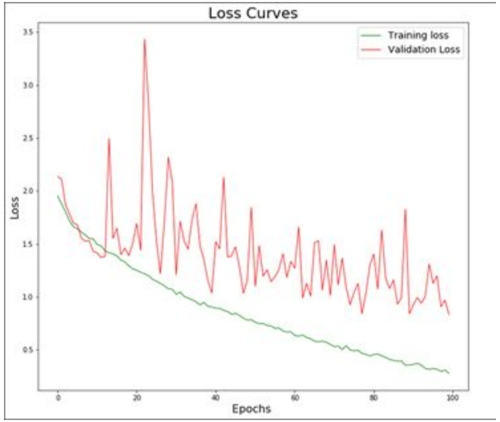
SVM classification pipeline:

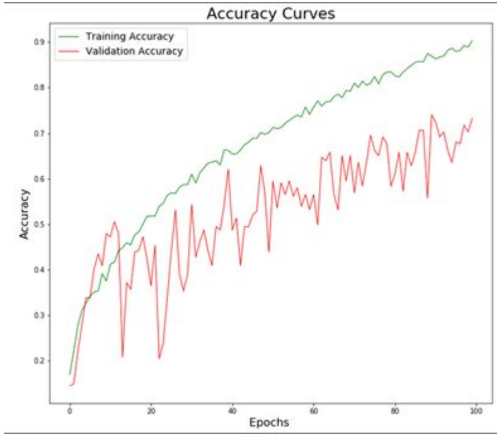
- Framing using a rolling window
- Apply Hamming filter
- Feature extraction
- Compute global statistics
- Train using best model from GridSearch hyperparameter tuning.
- Save best model and use to predict.

AUDIO ANALYSIS

Time Distributed CNN







TimeDistributed CNNs pipeline:

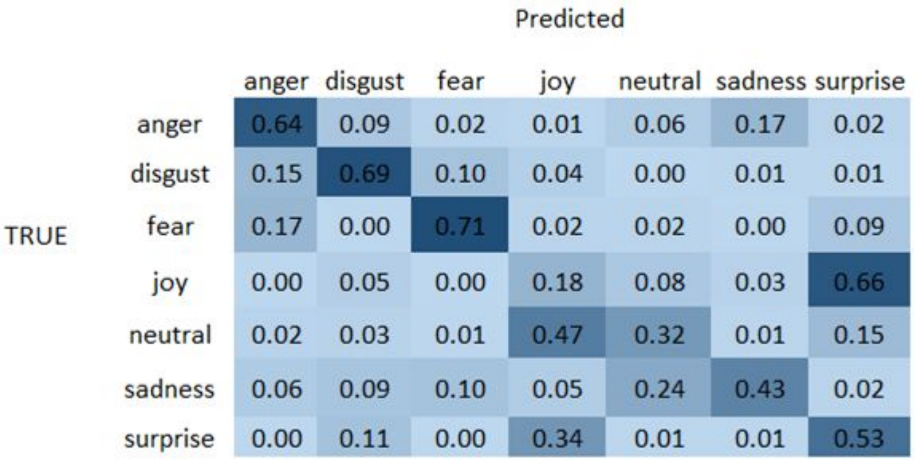
- Augment audio file to add noise.
- Log-mel-spectrogram extraction.
- Split spectrogram with a rolling window.
- Train model with original and augmented files.
- Save model and predict on test set.

AUDIO ANALYSIS

Experiment Results

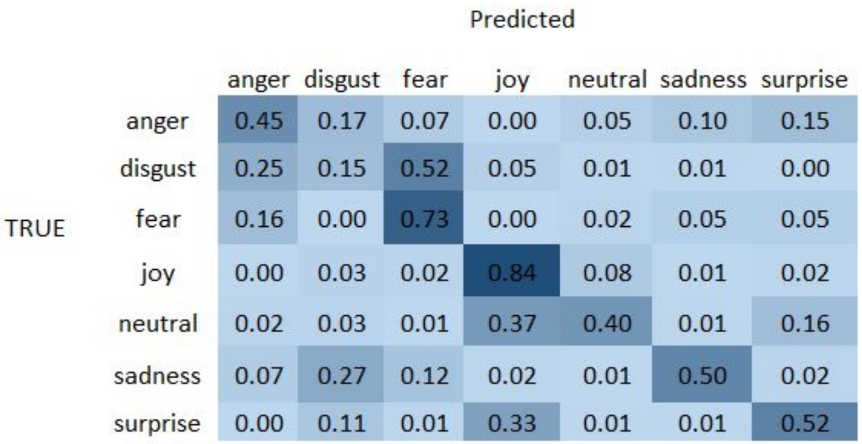
CNN performs better than SVM in some emotions reaching almost 84% in “Joy”. SVM may do better with a wider range of hyperparameters. SVM performs average overall in all classes.

Further experimenting with wider/smaller window and hop size may push accuracy for CNN much higher.



		Predicted						
		anger	disgust	fear	joy	neutral	sadness	surprise
TRUE	anger	0.64	0.09	0.02	0.01	0.06	0.17	0.02
	disgust	0.15	0.69	0.10	0.04	0.00	0.01	0.01
	fear	0.17	0.00	0.71	0.02	0.02	0.00	0.09
	joy	0.00	0.05	0.00	0.18	0.08	0.03	0.66
	neutral	0.02	0.03	0.01	0.47	0.32	0.01	0.15
	sadness	0.06	0.09	0.10	0.05	0.24	0.43	0.02
	surprise	0.00	0.11	0.00	0.34	0.01	0.01	0.53

SVM : C=3, gamma=0.005, kernel=rbf.



		Predicted						
		anger	disgust	fear	joy	neutral	sadness	surprise
TRUE	anger	0.45	0.17	0.07	0.00	0.05	0.10	0.15
	disgust	0.25	0.15	0.52	0.05	0.01	0.01	0.00
	fear	0.16	0.00	0.73	0.00	0.02	0.05	0.05
	joy	0.00	0.03	0.02	0.84	0.08	0.01	0.02
	neutral	0.02	0.03	0.01	0.37	0.40	0.01	0.16
	sadness	0.07	0.27	0.12	0.02	0.01	0.50	0.02
	surprise	0.00	0.11	0.01	0.33	0.01	0.01	0.52

CNN : Window Size= 128, Hop Size = 64

TEXT ANALYSIS

Textual analysis is used to predict emotion considering different properties of conversation - speaker classification, context awareness, etc

Emotion recognition goes beyond detecting conventional sentiments like:positive, negative or neutral feelings from text. The accuracy of NLP models depends very highly oncontext.

We intend to use:

- Bidirectional LSTM based Model
- Dialogue RNN models to analyse text.

Statistics	Train	Dev	Test
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.3	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003

LITERATURE REVIEW

- Emotion recognition goes beyond detecting conventional sentiments like: positive, negative or neutral feelings from text. The accuracy of NLP models depends very highly on context. We intend to use bc - LSTM based Model [6] and Dialogue RNN [7] models to analyse text.

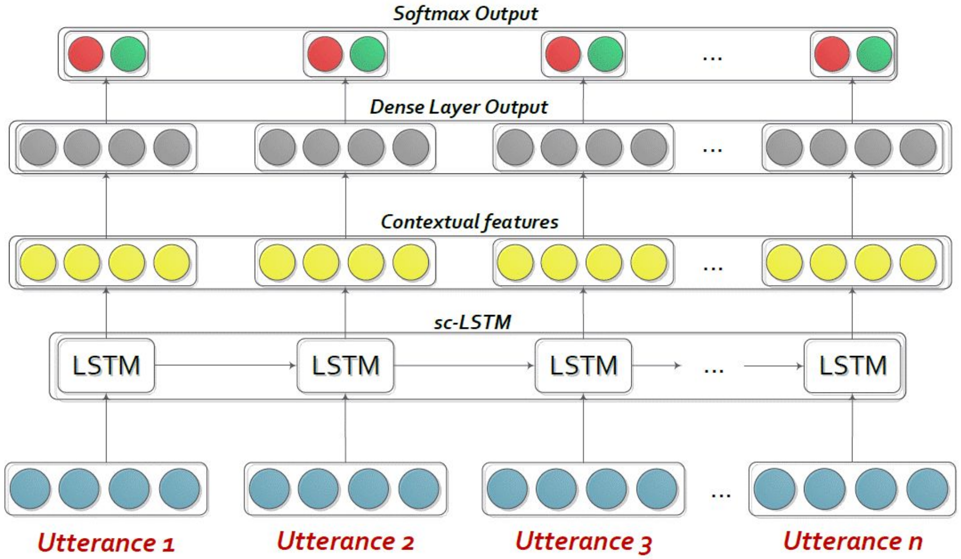
- Sukhtbatar has used memory networks in various NLP tasks including question - answering, machine translation, speech recognition and so on. Thus the memory networks were also successful in yielding state of art performance in emotion recognition using two memory networks for inter-speaker interaction. COGMEN uses Contextualized Graph Neural Network based models to show the importance of modeling information to model complex dependencies.

TEXT ANALYSIS

bc-LSTM model

Bidirectional LSTM model pipeline:

- Feature extraction using CNN and Glove embedding vector.
- Utterance passed into the BiLSTM layer
- Dense layer of a fully connected layer
- Using softmax layer for emotion classification
- Using categorical cross - entropy on each utterance.



TEXT ANALYSIS

Dialogue RNN

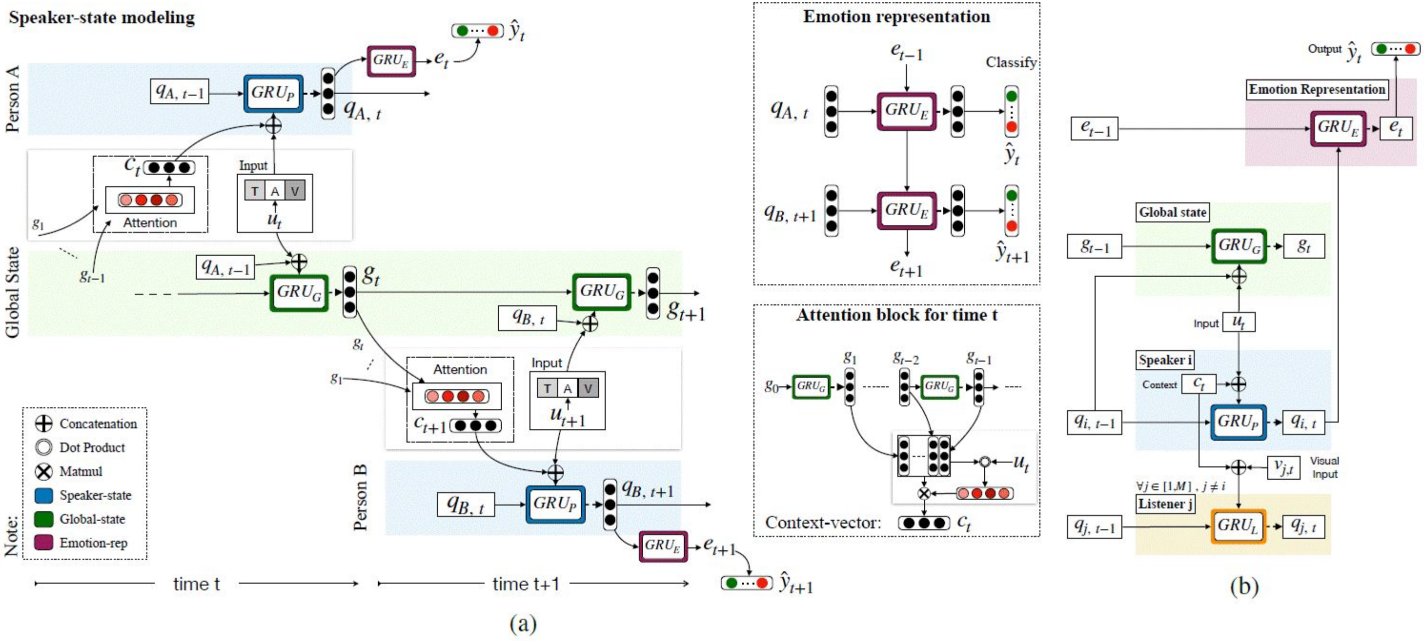
State of the art conversational emotion detection which models context by tracking individual speaker states throughout the conversation for emotion classification.

Models the emotion representation of the current utterance as a function of the emotion representation of the previous utterance and the state of the current speaker.

- Global state GRU - capture the context of a given utterance
- Party state GRU - keeps track of the state of individual speakers
- Speaker update GRU - frames the response based on the context, which is the preceding utterances in the conversation.
- Emotion GRU - emotionally relevant representation et of utterance from the speaker’s state and the emotion representation of the previous utterance

TEXT ANALYSIS

Dialogue RNN



TEXT ANALYSIS

Experimental Results

We used f1 score to evaluate the model on each emotion category and using its weighted average to compare the whole model.

- emotional classes disgust, fear and sadness are particularly poor due to the imbalance in the dataset which has fewer training instances.
- Dialogue RNN shows improved result due to its ability to differentiate the speakers in multispeaker context, unlike bc-LSTM

Models	Emotions							
	anger	disgust	fear	joy	neutral	sadness	surprise	w-avg
bc- LSTM	42.06	21.69	7.75	54.31	71.63	26.92	48.15	56.44
Dialogue RNN	40.59	2.04	8.93	50.27	75.75	24.19	49.38	57.03

VISUAL ANALYSIS

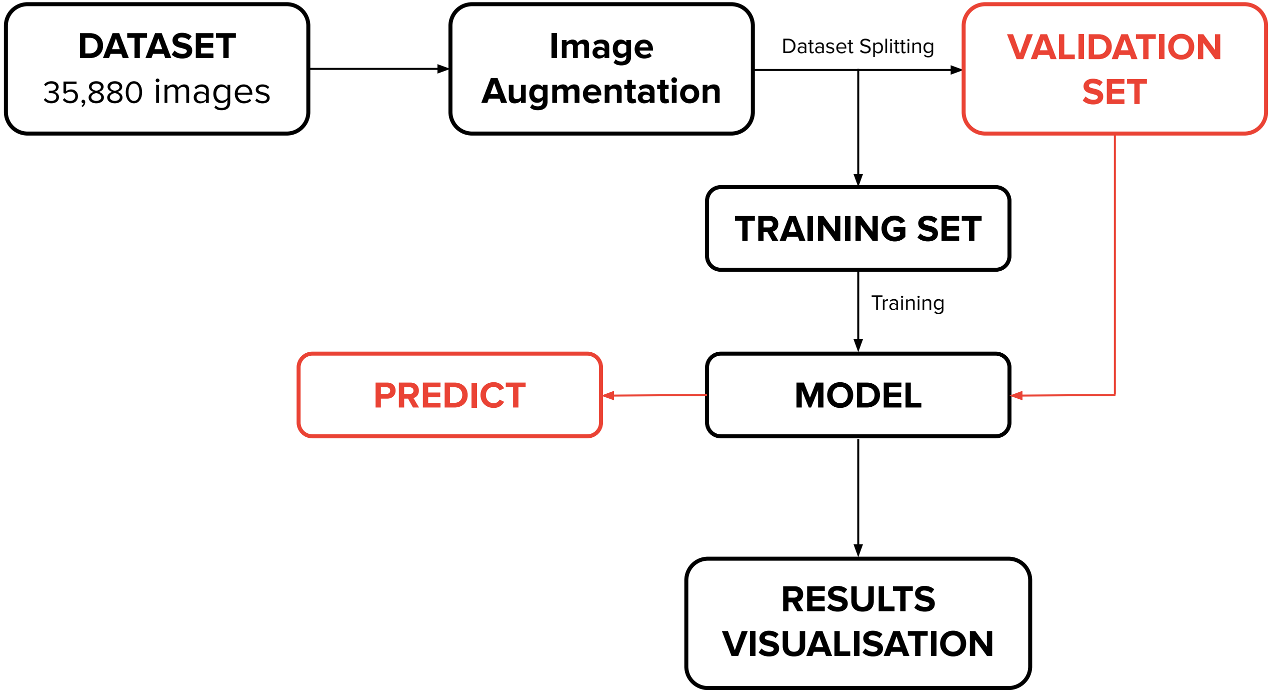
- Visual features prove very helpful in emotion detection as it involves features and facial expressions tracking which can be very distinctive for each emotion.
- A classifier must have effective features that are specifically tailored and optimized for its task if it is to produce accurate predictions.
- It should come as no surprise that mostly CNN's have performed well for classifying emotions, as shown by the fact that they are used in a variety of cutting-edge algorithms to perform this task.

LITERATURE REVIEW

- Hossain and Muhammad used the deep learning and cognitive wireless architecture to develop an audio-visual emotion detection system that could recognize patients' emotions in real time and automatically identify them in the context of Internet-based medical care. Through experiments, they assessed the system and established its value for the advancement of online medical care.

- To identify emotions in still images, Ng et al. combined CNN with transfer learning from ImageNet. The authors achieved 55.6% accuracy using the 2015 Emotion Recognition sub-challenge dataset of static facial expression.

VISUAL ANALYSIS

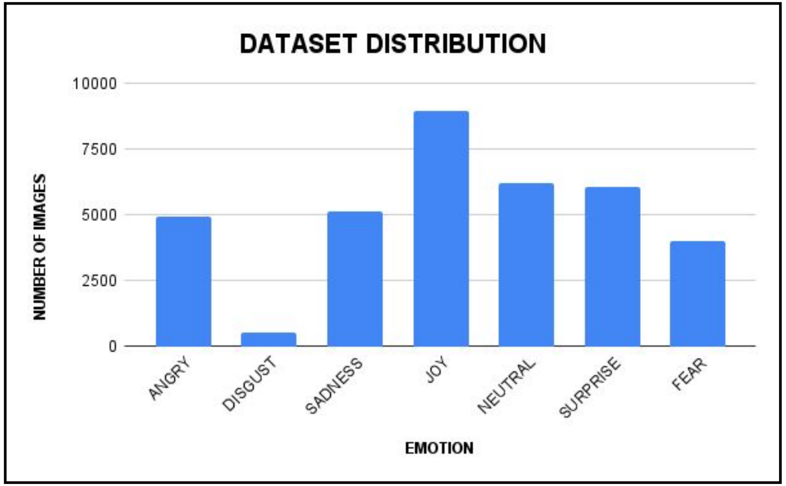


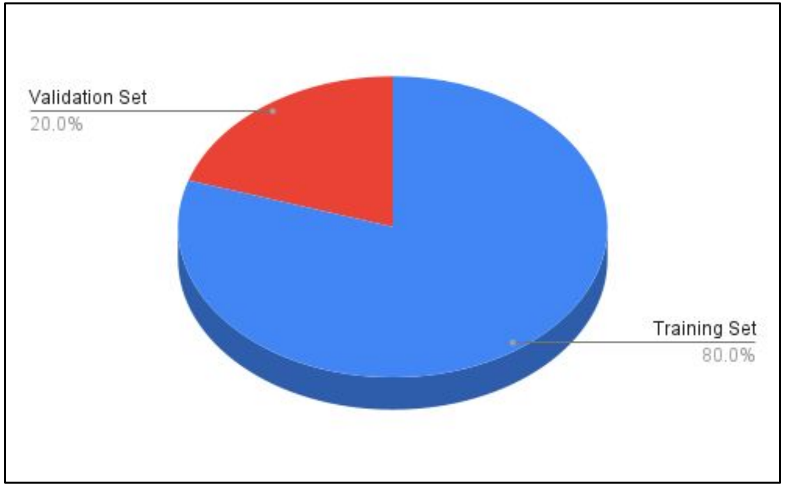
Proposed method for Visual Analysis

VISUAL ANALYSIS

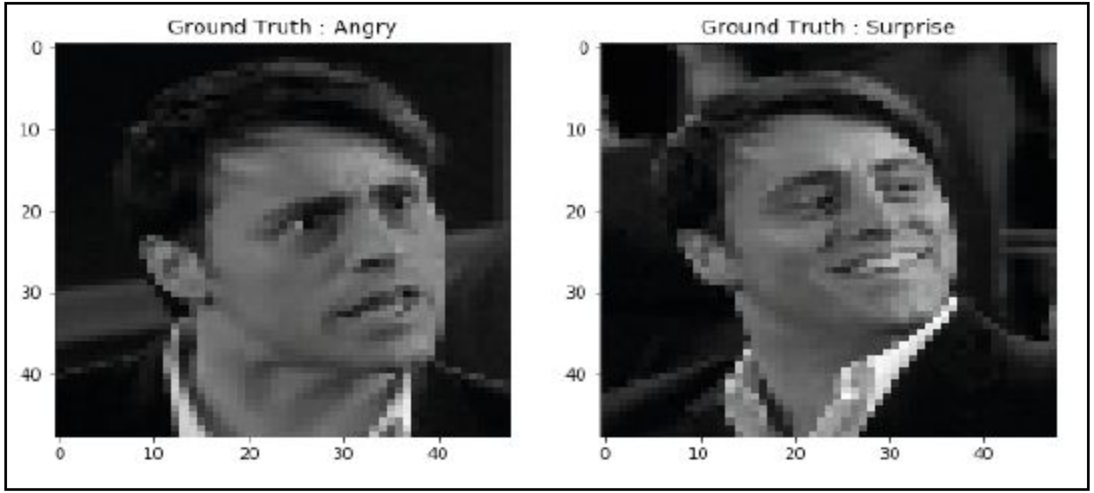
1. Dataset Preprocessing

The MELD contains a total of 35,880 images to which various augmentation techniques like scaling and flipping are applied and then the dataset is split into training and validation sets.





VISUAL ANALYSIS

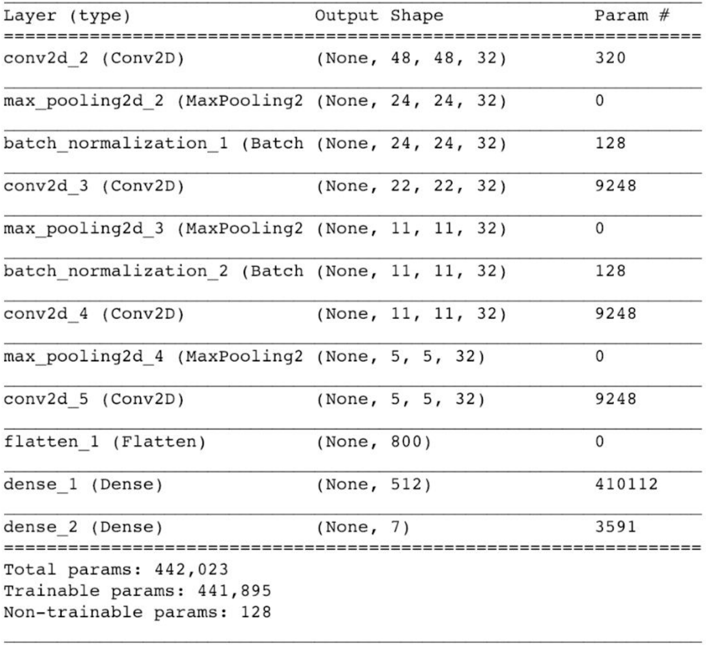


Example of a Frame from Dataset with ground truths

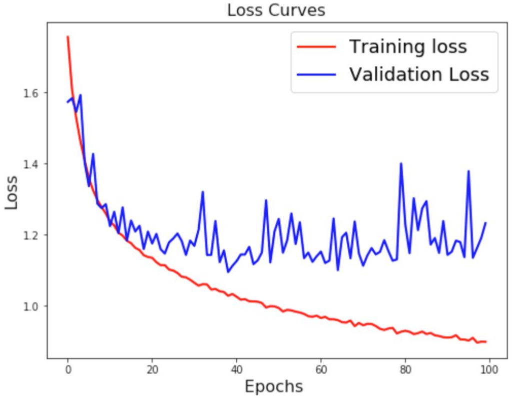
VISUAL ANALYSIS

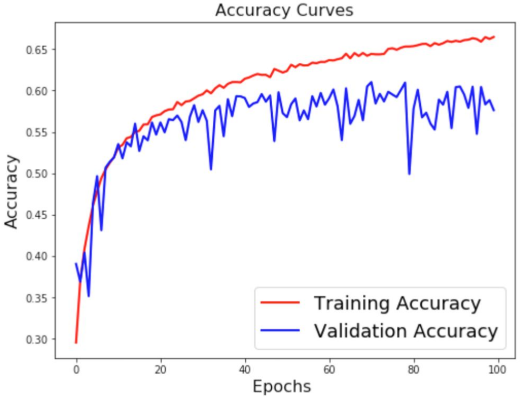
Model 1 :

CNN from Christopher and Martin’s paper "Facial Expression Recognition Using Convolutional Neural Networks: State of the Art” with optimizer RMSprop.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_2 (Conv2D)            (None, 48, 48, 32)        320
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 24, 24, 32)        0
_________________________________________________________________
batch_normalization_1 (Batch (None, 24, 24, 32)        128
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 22, 22, 32)        9248
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 11, 11, 32)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 11, 11, 32)        128
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 11, 11, 32)        9248
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 5, 5, 32)          0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 5, 5, 32)          9248
_________________________________________________________________
flatten_1 (Flatten)          (None, 800)               0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               410112
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 3591
=================================================================
Total params: 442,023
Trainable params: 441,895
Non-trainable params: 128
_________________________________________________________________
```

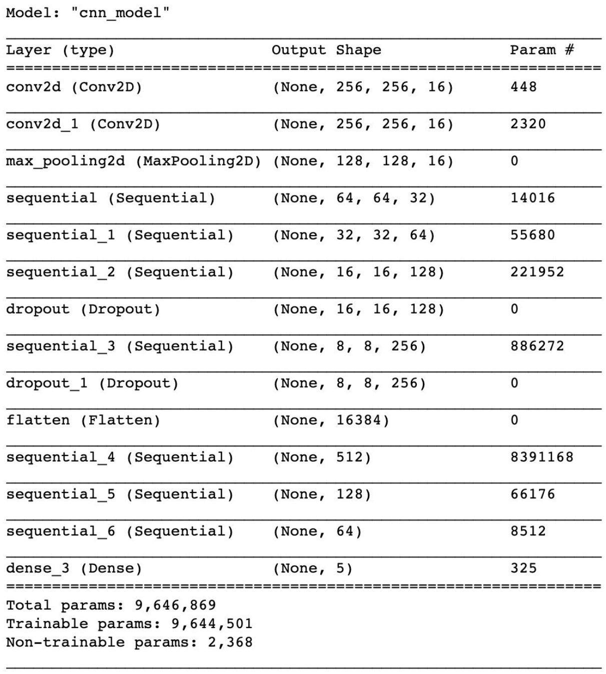




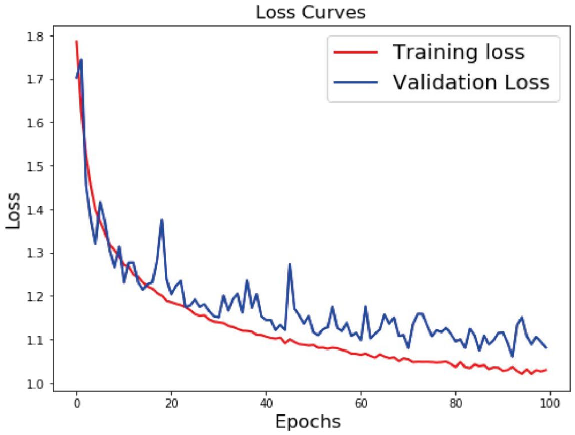
VISUAL ANALYSIS

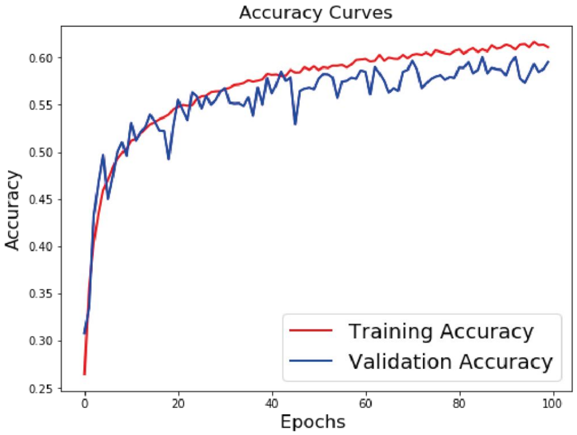
Model 2 :

Customized Convolutional Neural Network (CNN) model based on Inception-V3 Network with Adam optimizer.



```
Model: "cnn_model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 256, 256, 16)      448
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 256, 256, 16)      2320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 128, 128, 16)      0
_________________________________________________________________
sequential (Sequential)      (None, 64, 64, 32)        14016
_________________________________________________________________
sequential_1 (Sequential)    (None, 32, 32, 64)        55680
_________________________________________________________________
sequential_2 (Sequential)    (None, 16, 16, 128)       221952
_________________________________________________________________
dropout (Dropout)            (None, 16, 16, 128)       0
_________________________________________________________________
sequential_3 (Sequential)    (None, 8, 8, 256)         886272
_________________________________________________________________
dropout_1 (Dropout)          (None, 8, 8, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 16384)             0
_________________________________________________________________
sequential_4 (Sequential)    (None, 512)               8391168
_________________________________________________________________
sequential_5 (Sequential)    (None, 128)               66176
_________________________________________________________________
sequential_6 (Sequential)    (None, 64)                8512
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 325
=================================================================
Total params: 9,646,869
Trainable params: 9,644,501
Non-trainable params: 2,368
_________________________________________________________________
```





CONCLUSION AND FUTURE WORK

To conclude, detecting emotions require various context clues which can be present in either audio, video or simply text.

- To identify an appropriate methodology, we can see it is entirely data type dependent. For audio or speech, a CNN-LSTM works better than an SVM. However, some emotions are not distinct enough such as joy - surprise which can lead to false classification.
- For visual task, there is a need for a targeted data augmentation for a specific class as it is very difficult to predict a result for “disgust” class in MELD. Also in order to improve the model and approach state of the art results, we can try to include extracted facial features in design matrix using oriented gradients (Hog) on sliding windows.
- In context of Text analysis, Dialogue RNN proves to be a better model compared to LSTM, but the results can be improved by taking multi modal or ensemble approach. More information can be inferred from conversation using multi modality like emotion shifts in a conversation.

In the future we aim to do a conclusive study which will find an ensemble technique for multimodal emotion detection.

REFERENCES

Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," *2005 International Conference on Machine Learning and Cybernetics*, 2005, pp. 4898-4901 Vol. 8, doi: 10.1109/ICMLC.2005.1527805.

Wang, Jia-Ching & Wang, Jhing-Fa & Lin, Cai-Bei & Jian, Kun-Ting & Kuok, Wai-He. (2006). Content-Based Audio Classification Using Support Vector Machines and Independent Component Analysis. 4. 157-160. 10.1109/ICPR.2006.407.

Slimi, A., Nicolas, H. and Zrigui, M., 2022, July. Hybrid Time Distributed CNN-Transformer for Speech Emotion Recognition. In *Proceedings of the 17th International Conference on Software Technologies ICSOFT, Lisbon, Portugal* (pp. 11-13).

Issa Dias, Demirci F and Yazici A, 2020. Speech emotion recognition with deep convolutional neural networks In Biomedical Signal Processing and Control Vol. 59, doi: 10.1016/j.bspc.2020.101894.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R., MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. doi:10.48550/ARXIV.1810.02508

Bertero, D., & Fung, P. (2017). A first look into a Convolutional Neural Network for speech emotion detection. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5115-5119.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 873-883).

DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, E. Cambria, and G. Alexander. AAAI (2019), Honolulu, Hawaii, USA

THANK YOU