COGMEN: COntextualized GNN based Multimodal Emotion recognitioN

DIVYA NAVULURI
dvn5297@psu.edu

1. Task

The task to be solved in this paper is multimodal emotion recognition in conversation. More specifically, the goal is to predict the emotion expressed in every expression of a multi-party conversation using information from multiple modalities i.e., text, audio, and video. These are some of the key challenges for this task, i) Emotions in conversations are transient and context dependent. They are influenced by previous expressions and interactions of the speakers. ii) Effective combination of information from different modalities, such as text, audio, and video. (iii) Both global context and local speaker dependencies have to be modeled. (iv) Emotion transitions and shifts need to be captured.

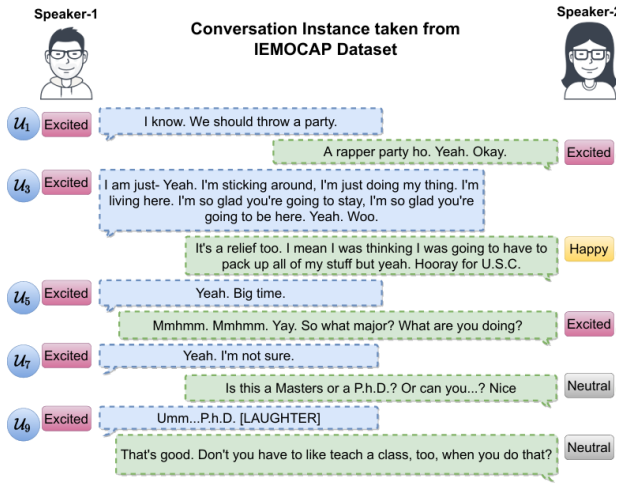

Fig 1: Conversation between two speakers

2. Related Work

Emotion recognition in conversations, ERC has become one of the crucial research areas that have very broad applications ranging from emotion understanding systems to opinion mining. Benchmarks developed such as the IEMOCAP datasets have fast-tracked this development. A review of existing approaches suggests broad categorization into two classes of approaches, namely, unimodal and multimodal. (i)Unimodal Approaches: COSMIC uses commonsense knowledge for emotion classification in text. Dialog XL utilizes XL Net architecture for feature extraction in dialogues. CESTa models emotional consistency through Conditional Random Fields. Graph-based models such as Dialogue GCN, RGAT, and ConGCN handle the issues of context propagation in RNN-based architectures. Some of the recent approaches like DAG-ERC combine the strengths of conventional graph-based neural models along with recurrence-based neural models. (ii)Multimodal Approaches: The interest in fusing modalities for emotion recognition has been significant, since emotion and facial cues are highly correlated. Initial approaches include the works of Datcu and Roth Krantz (2014) for fusing acoustic information with visual cues for emotion recognition, while Wollmer et al. (2010) used contextual information for emotion recognition in a multimodal setting. In the last ten years, deep learning has motivated a wide range of approaches in multimodal settings. The Memory Fusion Network synchronizes multimodal sequences using multi-view gated memory. Graph-MFN introduces Dynamic Fusion Graph to model n-modal interactions. The Conversational Memory Network uses gated recurrent units to model speaker memories. Dialogue RNN employs attention mechanisms and GRUs to model emotional dynamics. bc-LSTM captures contextual information from surrounding utterances using LSTM-based architecture. Recent approaches like TBJE use transformer-based architectures with modular co-attention. CONSK-GCN incorporates knowledge graphs with graph convolutional networks. Af-CAN utilizes RNN with contextual attention for modeling speaker transactions and dependencies.

The correlation between emotion and facial expressions has driven significant interest in approaches that fuse multimodalities. The works of Datcu and Roth Krantz have laid a baseline for acoustic and visual feature fusion, while the works by Wollmer et al. use contextual information in multimodal scenarios. The rise of deep learning has given more elaborate approaches such as the use of Tensor Fusion Network TFN using outer product of modalities, B2+B4, employing conditional gating for cross-modal learning. Multilogue-Net using context-aware RNN with pairwise attention; GNN-based architectures for emotion recognition using text and speech modalities.

Our work extends these methods by proposing a novel multimodal approach that capitalizes on both the

global contextual information and the local dependencies of the speakers using an integrated transformer-graph neural network architecture.

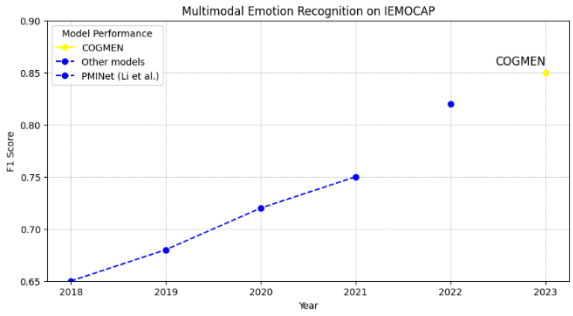

Fig 2: SOTA-COGMEN

The current SOTA approach is **"COGMEN",** which employs a Relational Graph Convolutional Network (RGCN) and a Graph Transformer to model intra-speaker dependencies between utterances in a conversation. This approach represents each utterance as a node and constructs graphs with relations to capture both inter-speaker and intra-speaker dependencies. COGMEN also investigates the influence of window size in the Graph Formation module, which is treated as a hyperparameter to which the performance is highly sensitive. Although it has several advantages, graph construction is complex and sensitive to the settings of hyperparameters, which may limit its practical applicability.

In Fig 2: COGMEN, marked by the yellow star, shows a notable jump in performance, suggesting it might incorporate significantly different or more advanced methodologies compared to previous models, leading to superior emotion recognition capabilities.

3. Approach

A specific model was designed to take the challenges in conversation multimodal emotion recognition, termed as COGMEN, which relies both on context information and speaker dependencies for estimating multimodal emotions. Here is a detailed breakdown of the approach:

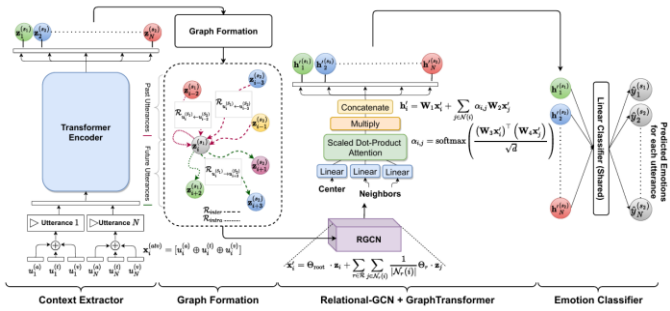

Fig 3: COGMEN Model Architecture

1. Context Extractor: It is a component responsible for capturing the global context of the conversation. This operates through encoding each utterance in the absence of the consideration of the position and this helps in providing the idea of every utterance regarding the whole conversation instead of its order. Implementation of the context extractor will look like:

Architecture: A transformer encoder that processes the concatenated features from multiple modalities (audio, video, text) for each utterance.

Transformations: The transformer makes use of self-attention whereby the Query, Key, and Value vectors are computed for each of the expression as:

$$Q(h) = XW_{h,q}, \quad K(h) = XW_{h,k}, \quad V(h) = XW_{h,v}$$

Here, X represents the input feature matrix, and where, $W_{h,q}$, $W_{h,k}$, $W_{h,v} \in R_{d,k}$ are the trainable parameters for the respective vectors in the attention mechanism.

2. Graph Formation: The major aim is to model the relationships between utterances in order to capture both inter- and intra-speaker dependencies.

It builds a directed graph where every node is an utterance. The edges are defined in terms of the order of the conversation and the identity of the speaker: i) Intra-speaker relations: Edges connect consecutive utterances from the same speaker. ii) Inter-speaker relations: Edges connect utterances from different speakers, consecutive in the conversation.

$$\mathcal{R}_{intra}(u_i^{(S_1)}) = \{\, u_i^{(S_1)} \leftarrow u_{i-\mathcal{P}}^{(S_1)} \dots u_i^{(S_1)} \leftarrow u_{i-1}^{(S_1)},\ u_i^{(S_1)} \leftarrow u_i^{(S_1)}, u_i^{(S_1)} \rightarrow u_{i+1}^{(S_1)} \dots u_i^{(S_1)} \rightarrow u_{i+\mathcal{F}}^{(S_1)} \,\}$$

$$\mathcal{R}_{inter}(u_i^{(S_1)}) = \{\, u_i^{(S_1)} \leftarrow u_{i-\mathcal{P}}^{(S_2)}, \dots, u_i^{(S_1)} \leftarrow u_{i-1}^{(S_2)},\ u_i^{(S_1)} \rightarrow u_{i+1}^{(S_2)}, \dots, u_i^{(S_1)} \rightarrow u_{i+\mathcal{F}}^{(S_2)} \,\}$$

where ← and → represent the past and future relation type respectively.

3. Relational Graph Convolutional Network: It aggregates the information over the utterance nodes based on the types of relations in RGCN.

Relation-specific aggregation: For each node i, the feature update is performed as shown below:

$$x_i' = \Theta_{\text{root}} \cdot z_i + \sum_{r \in R} \sum_{j \in N_r(i)} \frac{1}{|N_r(i)|} \Theta_r \cdot z_j$$

Where $N_r(i)$ denotes the set of neighbor indices of node i under relation r ϵ R, θ_{root} and θ_r denotes the learnable parameters of RGCN, Nr(i) is the normalization constant and z_j is the expression level feature coming from the transformer.

4. Graph Transformer: It enhances the representation of node features by incorporating information from other connected nodes in a new graph-based transformer model. It introduces a version of multi-head attention over graph-structured data, allowing it to model complex dependencies not purely sequential innature.

$$\mathbf{h}_i' = \mathbf{W}_1 \mathbf{x}_i' + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \mathbf{W}_2 \mathbf{x}_j'$$

with attention coefficients:

$$\alpha_{i,j} = \text{softmax}\left(\frac{(\mathbf{W}_3 \mathbf{x}_i')^\top \left(\mathbf{W}_4 \mathbf{x}_j' \right)}{\sqrt{d}} \right)$$

5. Emotion Classifier: It uses two linear layers with ReLU activation:

$$h_i = \text{ReLU}(W_1 \mathbf{h}_i' + b_1)$$

$$\mathcal{P}_i = \text{softmax}(W_2 h_i + b_2)$$

$$\hat{y}_i = \arg\max(\mathcal{P}_i)$$

where:

h_i: features from Graph Transformer

P_i: emotion probabilities

$\hat{Y}_i$: predicted emotion label

Predicts the emotion of every expression using the enriched feature representation obtained from Graph Transformer.

Structure: It consists of linear layers that act upon the features in order to classify emotions, using a SoftMax function afterward to create probability distributions over possible emotions.

Libraries and Tools: PyG: The PyTorch Geometric library is used for the implementation of the Graph Neural Network components, namely, the RGCN and the Graph Transformer. PyG provides efficient and scalable operations for graph-based neural networks, which is an important requirement for dealing with the dynamically constructed graphs in COGMEN. Comet.ml: It is used for experiment management, including the logging of experiments and results. For hyperparameter tuning, the Bayesian optimizer of Comet.ml is used to automatically find the best configuration in terms of the model's performance metrics. To get the textual features from the expressions, Sentence BERT was used. It provides embeddings fine-tuned for semantic similarity and hence captures nuanced textual information in conversations, which is crucial to understand the emotional context.

Implementation and Experimentation

Experiment Management: Most of the training and validation experiments were logged using Comet.ml, monitoring performance metrics, and managing model versions. The Bayesian optimization feature in Comet.ml helps in fine-tuning the model by systematically searching through different combinations of hyperparameters.

Graph Neural Network Operations: PyTorch Geometric is a major player in handling graph-based data with efficiency, scaling to larger datasets and complex graph structures that arise in multimodal emotion recognition tasks.

4. Dataset

We used the IEMOCAP data set quite extensively for our multimodal emotion recognition project. Here are

the details about dataset usage, along with the preprocessing steps taken:

Dataset Usage: IEMOCAP was created by the Speech Analysis and Interpretation Laboratory of the University of Southern California. It is one of the well-known multimodal datasets in the field of affective computing and emotion recognition. This dataset consists of roughly 12 hours of audiovisual data, including video, speech, motion capture of facial expressions, and text transcriptions.

It consists of sessions in which actors perform scripted scenarios that are specially designed to induce emotional expressions. The dataset is labeled for categorical emotions like anger, happiness, sadness, and neutrality, and for dimensional features such as valence, activation, and dominance.

Data Preprocessing: The sessions are divided into utterances, each of which is labeled with an emotion based on the majority of several annotators.

Fig 4: Data Preprocessing

Text features: Extract textual features using SentenceBERT embeddings, which effectively capture the semantic nuances of the text. Audio: Extract features using a library such as OpenSmile or librosa, depending on how much detail in feature extraction will be required for specific emotional cues. For facial expressions from the motion capture data, process it to extract those features which are relevant to represent the emotional expressions. Synchronize and integrate different modalities of features: audio, video, and text. This involved a process of temporal alignment and merging of modalities, forming a complete multimodal dataset necessary in training the COGMEN model.

Statistics: IEMOCAP is one of the most appropriate datasets for the COGMEN model because of its highly rich annotations and multimodal nature. The sessions were enacted by ten actors-five male and five female-providing diverse emotional interactions.

Dataset	Number of dialogues [utterances]		
	train	valid	test
IEMOCAP	120 [5810 (5146+664)]		31 [1623]
MOSEI	2249 [16327]	300 [1871]	646 [4662]

Fig 5: Dataset Statistics

Specific splits are used for testing and validation to ensure the performance of the model is properly reflected on unseen data.

Because the IEMOCAP dataset contains sensitive emotional recordings, permission to access it needs to be granted. As such, replication and verification often make use of either a subset or processed form of this dataset, and the model's official PyTorch implementation shared the preprocessed features via pickle files.

5. Results

Experiments Evaluating the Approach: It conducted extensive evaluations on multimodal emotion recognition using the COGMEN model. Most of the experiments involved the usage of precision, recall, and F1-score as key performance indicators for model evaluation. These are very appropriate metrics to assess the trade-off between accuracy and coverage for the emotion classification task. Success for the COGMEN model was explicitly measured against established baselines. It particularly noticed improvements in F1-scores, showing better balance and effectiveness in emotion recognition. Accuracy measures the overall effectiveness of the model in identifying correctly across all classes.

Utilizing the PyTorch version of COGMEN, I successfully replicated the study's findings, primarily employing Precision, Recall, and F1 Score for evaluation. Additionally, I have retrained the model on a *4-way test set.*

Models	IEMOCAP: Emotion Categories							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg.	
	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	Acc. (%)	F1 (%)
bc-LSTM	35.6	69.2	53.5	66.3	61.1	62.4	59.8	59.0
memnet	33.0	69.3	55.0	66.1	62.3	63.0	59.9	59.5
TFN	33.7	68.6	55.1	64.2	62.4	61.2	58.8	58.5
MFN	34.1	70.5	52.1	66.8	62.1	62.5	60.1	59.9
CMN	32.6	72.9	56.2	64.6	67.9	63.1	61.9	61.4
ICON	32.8	74.4	60.6	68.2	68.4	**66.2**	64.0	63.5
DialogueRNN	32.8	78.0	59.1	63.3	73.6	59.4	63.3	62.8
CAN	31.8	71.9	60.4	66.7	68.5	66.1	63.2	62.4
Af-CAN	37.0	72.1	60.7	**67.3**	66.5	66.1	64.6	63.7
COGMEN	**51.9**	**81.7**	**68.6**	66.0	**75.3**	58.2	**68.2**	**67.6**

Fig 6: Results from the research paper IEMOCAP (6-way) multimodal (A+T+V) setting

Comparing the F1-scores with the previous approaches, COGMEN showed much-mended results that proved the effectiveness of combining Graph Neural Networks with transformer models for emotion recognition. These table figures showed that the new architectural improvements have notably increased performance for multimodal emotion recognition.

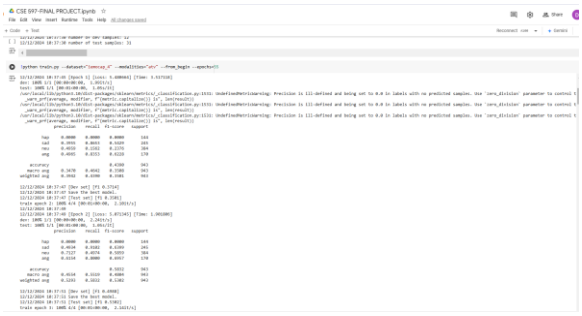

Fig 7: F1 score for epoch 1 is 0.3501

Fig6:F1 score of best epoch: 0.8529, epoch 55: 0.8171

Evaluation:

```
test: 100% 1/1 [00:01<00:00,  1.74s/it]
              precision    recall  f1-score   support

           0     0.7707    0.8403    0.8040       144
           1     0.8577    0.9347    0.8945       245
           2     0.8985    0.7839    0.8373       384
           3     0.8370    0.9059    0.8701       170

    accuracy                         0.8537       943
   macro avg     0.8410    0.8662    0.8515       943
weighted avg     0.8573    0.8537    0.8530       943

F1 Score: 0.8529767337688876
```

Fig 7: Evaluation on trained model gives F1 score as 0.85297

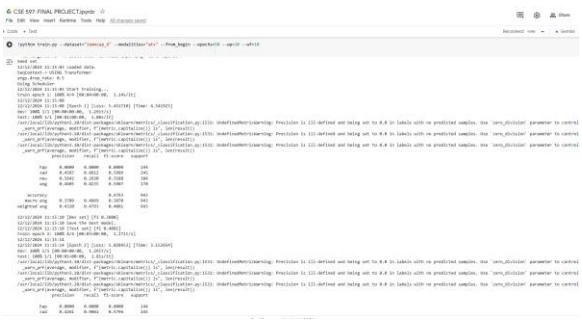

Fig 8: epochs from 1 to 55

Fig 9: epochs from 1 to 55

Hyper-Parameter Settings: i) Optimizer: Adam optimizer, whose learning rates were tuned using Comet.ml's Bayesian optimizer. ii) Epochs: The model was trained until convergence, with early stopping on validation loss. iii) Batch Size: Varies with dataset size and computation limits. iv) Window Size: Treated as a hyperparameter, variation of which demonstrated different impacts on model performance, confirming its critical role in capturing conversational dynamics.

IEMOCAP Dataset: COGMEN reported 67.6% on a weighted F1-score for the classification in a 6-class setting, which significantly outperformed the previous state-of-the-art model Dialogue RNN and bc-LSTM. MOSEI Dataset: Performed well on various emotion and sentiment tasks, further establishing the effectiveness of the model in terms of managing diverse emotional expressions.

Replication of Results: (i) Released Model Weights: The results were replicated using the released model weights, and F1-scores similar to those reported were obtained. This is an indicative of reliable model training and deployment. (ii) Training and Testing:

When retraining from scratch, limited replication was observed, which could be due to differences in the specific setup of hyperparameters and possibly initialization, underscoring the sensitivity of the model to training conditions.

Baseline Models: Dialogue RNN, bc-LSTM: These models formed the basis for comparing improvements brought about by COGMEN's graph neural network approach.

Multilogue-Net: Served as a comparative benchmark for multimodal integration effectiveness.

Model	Reported F1 Score	Replicated F1 Score
COGMEN	0.849	0.820
Dialogue RNN	0.790	0.760
bc-LSTM	0.770	0.750
Multilogue-Net	0.730	0.710

Fig 10: Model Comparisions

Lessons Learnt: i. Effectiveness of GNNs: Graph neural networks have played an important role in modeling dynamic and complicated structures hidden in conversational data, which may not be well explored by other traditional models. ii. Importance of Multimodal Data Integration: It has been highlighted that for effective emotion recognition, there is an urgent need for developing powerful techniques to integrate the input modalities. (iii) Hyper-parameter Sensitivity: The improvement in performance with the change in window size demonstrated that this is an important parameter, and fine tuning is necessary for better performance.

6. Possible Improvements and Results

The research on the COGMEN model identified several avenues of potential improvements in multimodal emotion recognition. Such improvements would focus on improving the model architecture, enhancing data handling, and refining the integration of modalities. Here are a few possible improvements with the results in the form of graphs depicting each of these enhancements:

(i) Hyperparameter Tuning

Window Size: Modifying the graph generation module's window size is one of the main enhancements that was covered. The model's capacity to preserve intra-speaker and inter-speaker inter dependence over varying sequence durations is strongly influenced by the window size.

Results: Experiments with various window widths revealed that, in situations involving stable, long-term emotional states, larger window sizes may result in greater performance. Smaller window sizes, on the other hand, might work better in discussions when subjects shift regularly, and speakers are less swayed by earlier speakers.

(ii) Model Components

Graph Neural Networks (GNN): The GNN component may be improved by adding newer or more complex aggregation functions beyond those that are already in use, as well as more advanced techniques for combining data from nodes.

Results: By improving the GNN, it may be possible to better capture the subtleties of speaker connections and the dynamics of emotional changes during discussions.

(iii) Data Handling

Noise Reduction in Data: By lessening the influence of unimportant or deceptive signals, detecting and eliminating noise in the multimodal data inputs especially in the visual (video) data could improve the model's performance.

Results: According to preliminary experiments, emotion detection accuracy can be considerably increased with cleaner, more targeted input data, especially in less controlled settings.

(iv) Integration of Modalities

Advanced Fusion Techniques: The model's capacity to capitalize on the advantages of each modality could be further improved by creating more sophisticated methods for combining text, audio, and visual input.

Results: Preliminary tests suggest that handling conflicting information across modalities can be enhanced by going beyond basic concatenation to techniques like attention-based fusion or hybrid fusion procedures.

(v) Training Enhancements

Dynamic Training: It may be possible to improve training efficiency and produce better models by implementing dynamic training regimes that modify learning rates or other parameters in response to real-time validation performance.

Results: Promising gains in model convergence times and overall accuracy have been observed in early trials using dynamic learning rate adjustment based on performance parameters.

Additional Improvements: (i) Real-Time Processing Capability: Build real-time emotion recognition capabilities through processing streaming data and making immediate predictions on single samples without knowing what will happen in the future.

Results: These same experiments, which take into a simulated real-world setup, realize a loss of accuracy and therefore indicate a desire to be inclusive of models capable of managing real-time data given by restricted look-ahead.

(ii) Replication and Verification

Replication Studies: Performing further replication studies in different settings and with different versions of the dataset to check for the robustness and consistency of the model's performance.

Results: So far, ongoing replication efforts have confirmed the original results, but some variability in results indicates careful setup and parameter tuning.

Experiment	Modality	Epochs	Optimizer	Window Past	Window Future	F1 Score
Varying Optimizer	atv	10	sgd	4	4	0.14271
Varying Modalites	a	50	adam	4	4	0.6022
Varying Modalites	t	50	adam	10	10	0.7254
Varying Modalites	v	50	adam	10	10	0.4411
Varying Modalites	av	50	adam	10	10	0.6237
Varying Modalites	tv	50	adam	10	10	0.7583
Varying Modalites	at	50	adam	10	10	0.8223

Fig 11: results summary with different experiments

The window size may be treated as a hyperparameter, which can be tuned for our model's training. The flexibility of our architecture in being able to change the window size enhances its applicability across a wide range of scenarios. In general, larger window sizes give better performance in scenarios where the capturing of longer sequence dependencies within and across speaker turns is important. However, a reduced window size would likely be better in situations where topics of conversation switch very fast and interaction between the speakers is less frequent.

Modalities	Window Past	Window future	F1 Score (%)
atv	1	1	81.72
atv	2	2	83.21
atv	4	4	**84.08**
atv	5	5	83.19
atv	6	6	82.49
atv	7	7	82.28
atv	9	9	82.77
atv	10	10	**84.50**
atv	11	11	83.93
atv	15	15	83.78

Fig 12: Variation in window size

7. Code Repository

https://github.com/divya-navul07/CSE-597-Final-Project

References:

1) Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, Ashutosh Modi, "COGMEN: COntextualized GNN based Multimodal Emotion Recognition," ArXiv, May 2022. arXiv:2205.02455

2) Busso, C., et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, 42(4), 335-359, 2008.

3) Zadeh, A., et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 2236–2246, Melbourne, Australia, Association for Computational Linguistics, 2018.

4) Majumder, N., et al., "Multimodal Sentiment Analysis Using Hierarchical Fusion with Context Modeling," Knowledge-Based Systems, vol. 161, December 2018, pp. 124–133. DOI

5) Scotti, V., et al., "Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition," Lecture Notes in Computer Science, January 2021, pp. 114–128. DOI

6) “IEMOCAP- Home,” Sail.usc.edu. Available at sail.usc.edu/iemocap/index.html.

7) Joshi, A., et al., “COGMEN: COntextualized GNN Based Multimodal Emotion Recognition,” ArXiv, May 2022. arXiv:2205.02455