Emotion Recognition in Dialogue Systems

Mehak Piplani
Disha Kedige Chandrashekarachar
Rhushabh Vaghela

1 INTRODUCTION

Conversations are complex since they are governed by various variables like a topic, speaker's viewpoint, speaker's personality, argumentation logic, intent, etc [1] that affect the flow of a conversation and the emotional dynamics of the participants.

During a conversation, individual utterances are affected by the mental state, intent, and emotional state of the participants. Although the emotional state of these speakers cannot be observed directly, it can be inferred from the utterances through observable variables such as facial expressions, gestures, pitch, and acoustic indicators [2].

Inspired by the study [3], this work is focused on studying AI's capability to understand human emotions in real-time. Working with personal assistants such as Cortana, Alexa, Siri in our day-to-day life, we believe that identifying emotions in conversations is a core step towards fine-grained conversation understandings. It will allow these agents to adapt according to the responses and behavioral patterns. We aim to identify various verbal and non-verbal cues and their impact in determining the emotion of a person during conversations. Several strong multi-modal baselines are also proposed to solve the problem of emotion recognition in conversation.

2 LITERATURE REVIEW

Emotion recognition in conversation (ERC) has become an active area of research in recent years. The main approach towards ERC is contextual modeling utilizing deep-learning algorithms. The study [4] used recurrent neural networks for multi-modal ERC followed by CMN [5] and ICON [5] both utilize gated recurrent unit (GRU) and memory networks. Majumder et al. [6] proposed a recurrent network to model the party-state, global state, and emotional dynamics. Ghosal et al. [7] proposed a graph neural network-based model to encode speaker dependencies and temporal information. Then, Sahay et al. [8] used Relational Tensor Network, a new method to fuse the modalities showing an improvement over the early fusion baseline. Recently, the study [9] proposed a context-aware RNN, Multilogue-Net for ERC.

3 DATASET AND STATISTICAL ANALYSIS

We are utilizing the dataset: Multi-modal EmotionLines Dataset (MELD) [10]. It has more than 1400 dialogues and 13000 utterances from the Friends TV series. The dataset is divided into three sets: Train, Development, and test with more detailed statistics mentioned in Table 1. An example from the dataset showcasing an emotion shift from surprise to sadness in a single conversation is visualized in Fig 2 from [10]. The Fig 2 also defines how the emotional dynamics depend on both the previous utterances and their associated emotions.

Multiple characters (aged 25 to 32 yrs) enacted in the series with six main characters as visualized in Fig 3b. Each utterance in dialogue has been labeled by emotion from Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear as illustrated in Fig 1. A character-wise distribution of the emotion labels can be visualized in Fig 3a. MELD also has sentiment (positive, negative, and neutral) annotation for each utterance as described in Fig 3c.

Statistics	Train	Dev	Test
# of unique words	10,643	2,384	4,361
Avg.utterance length	8.0	7.0	8.2
Max utterance length	69	37	45
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Table 1: Dataset statistics

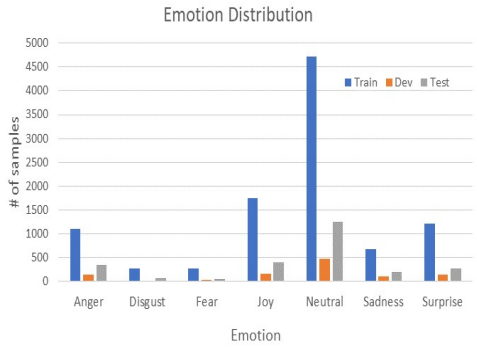


Figure 1: Emotion distribution for three sets of the dataset: train, development, and test.

CSCI 535, May 07, 2021, Los Angeles, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

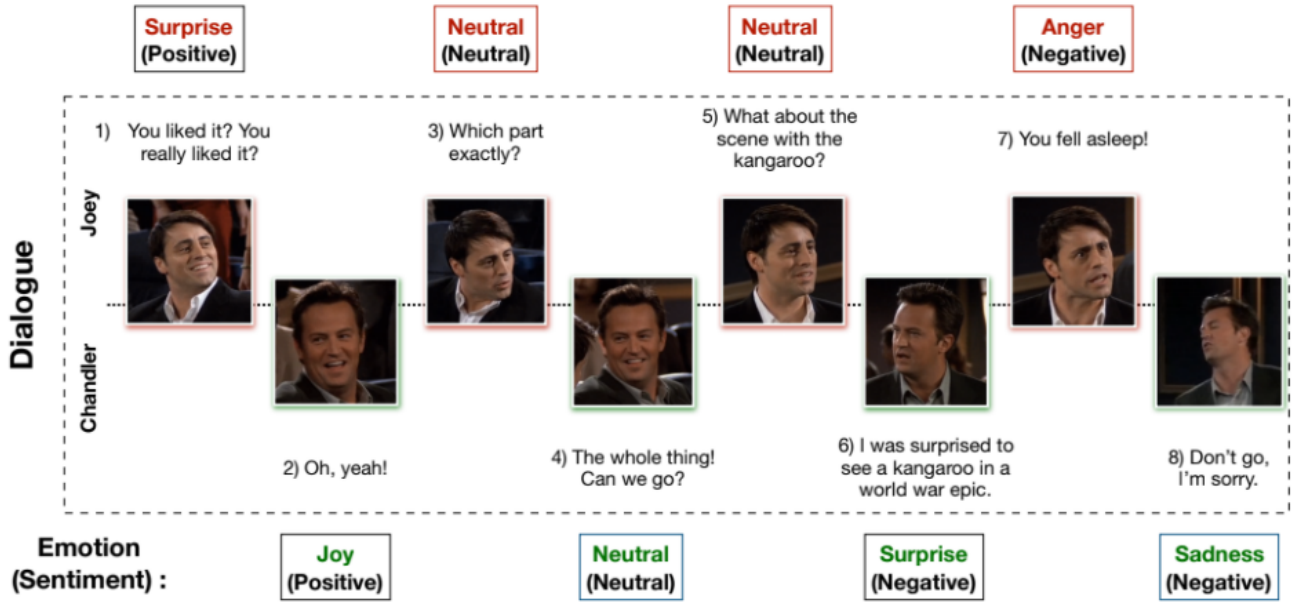
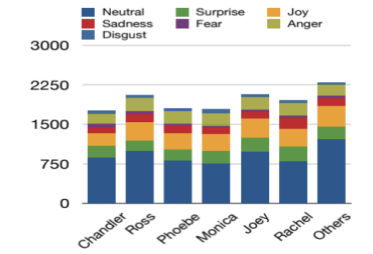
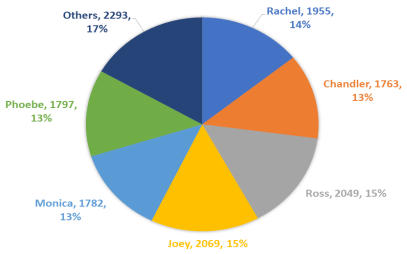


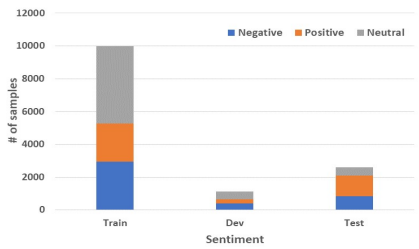
Figure 2: An example from the dataset showing an emotion shift of speakers in a dialogue from [10].



(a) Emotion distribution character wise [10]



(b) Character distribution



(c) Sentiment distribution

Figure 3: Dataset statistics

4 FEATURE EXTRACTION

4.1 Acoustic Features

The popular toolkit open smile [11] has been utilized to extract acoustic features. The toolkit extracts 6373-dimensional embeddings constituting several low-level descriptors and various statistical functionals of varied vocal and prosodic features. Two different types of acoustic feature representations were considered. For the first, an L2-based feature selection (SVM) was employed to get a dense representation of the overall audio segment, and for the second only the functionals (Means, Centroid, Peaks, Segments, Discrete Cosine Transformation (DCT), Zero Crossings, etc) were considered. These extracted features capture different characteristics of the human voice and have been shown related to emotions [12].

4.2 Lexical Features

The utterance transcripts were tokenized to extract 300-dimensional embeddings using the pre-trained Glove model [13] and then fed to a 1D-CNN to extract 100-dimensional textual features as demonstrated by Poria et al. [4] Another set of embeddings were extracted using a method called BERT (Bidirectional Encoder Representations from Transformers) [14]. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning both left and right context in all layers [14]. As a result, 768-dimensional features were created.

4.3 Visual Features

For the visual features, each video was divided into frames. Each frame was passed through a ResNet model [15] to extract its embedding. Later, average-pooling was used to merge the embedding for each frame corresponding to one training sample. To deal with high dimensional challenge from a massive amount of video data

and extract effective features by self-learning model based on deep learning, a short term memory (LSTM) [16] auto-encoder was experimented with. First, each frame was pre-processed (resizing, normalizing and augmentations like random rotation, shifts, shear and flips) and then passed through an LSTM encoder network embeddings to extract a 4096-dimensional feature vector.

5 METHODOLOGY

5.1 Data Pre-processing

To capture an efficient relationship between the features and their labels in this skewed dataset, techniques like Up sampling, Down sampling, and weighted loss (class balanced [17]) were utilized. The training samples belonging to class "neutral" and "Joy" were down-sampled to maintain a constant ratio of examples between all the class labels. Various experiments and learnings from the article [18] led to the belief that keeping the training samples belonging to class "disgust" and "fear" was affecting the efficiency of the model and hence only five labels are being considered for further experiments and observations.

5.2 Uni-modal and Bi-modal Networks

The motive of our experiments was to understand which modality captures the context efficiently and how these modalities interact with each other. To get an idea of the effect of each modality on the emotion label, we developed uni-modal models using Multi-Layer Perceptron (MLP) networks consisting of two or three linear layers followed by RELU activation. The idea behind experimenting with an MLP model was to form a baseline and observe how a simple model works for our problem.

Further, to experiment how each of these modalities comprehends each other, bi-modal networks were considered. The bi-modal networks combine modalities pairwise via the techniques of early fusion and late fusion. For early fusion, features were concatenated and modeled through an MLP network, and for late fusion, a weighted average of the individual model probabilities was considered for label prediction.

5.3 Multi-modal Networks

Fig 4 showcases instances from the dataset that demonstrate the importance of multi-modal cues. For example, the utterance "Great, now he is waving back" has its actual emotion classified as Disgust but if we consider only the text modality, we might interpret it wrong as Joy. Hence, it is important to focus on all the modalities simultaneously.

Emotion for a dialogue predominantly depends on four factors interlocutor state, interlocutor intent, the preceding and future emotion, and the context of the conversation [9]. To capture the above-mentioned factors and model the interactions between all the three modalities, the following current state-of-the-art methods were experimented with: Tensor fusion networks inspired from [8] and Dialogue RNN inspired from [6].

MLP Networks: Multi-modal networks were created via the techniques of early fusion and late fusion for MLP networks consisting of two or three linear layers followed by RELU activation. For early fusion, features were concatenated and modeled through an MLP network, and for late fusion, a weighted average of the individual model probabilities was considered for label prediction.

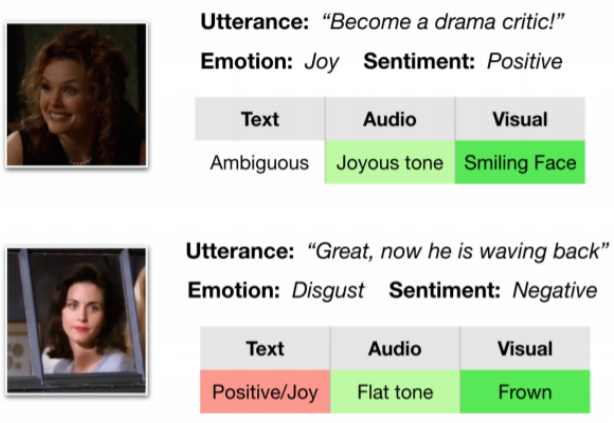


Figure 4: Importance of multi-modal cues. Green shows primary modalities responsible for sentiment and emotion from [10].

CNN-LSTM Networks: To incorporate context of an utterance, a sequence network is required. As inspired from the study [19], networks comprising of convolution layers followed by Long short-term memory (LSTM) layers were built. These networks aim to learn local and global emotion-related features. The techniques of early fusion and late fusion were incorporated for this setup similar to the MLP networks.

Tensor Fusion Network: The Tensor Fusion Network (TFN) was chosen to understand how to incorporate different modality interactions and how they affect emotion label prediction. This model learns both the intra-modality and inter-modality dynamics end-to-end as visualized in Fig 5. Inter-modality dynamics are modeled with a multi-modal fusion approach, named Tensor Fusion, which explicitly aggregates uni-modal, bi-modal, and multi-modal interactions. The intra-modality dynamics are modeled through Embedding sub-networks for lexical, visual, and acoustic modalities, respectively. These sub-networks are composed of Convolution layers followed by Bidirectional LSTM layers. Then, the output tensors from each sub-network are fused by an outer product forming the fusion layer having no learnable parameters. The final component of the network is an inference sub-network made of linear layers for emotion inference.

Dialogue RNN: The Dialogue RNN network has an effective mechanism to model context by tracking individual speaker states throughout the conversation for emotion classification. It employs three stages of gated recurrent units (GRU) as visualized in Fig 6 to model emotional context in conversations. The spoken utterances are fed into two GRUs: global and party GRU to update the context and speaker state, respectively. In each turn, the party GRU updates its state ($Q_{p,t}$ representing state for participant p at time t) based on the utterance spoken, the speaker's previous state, and

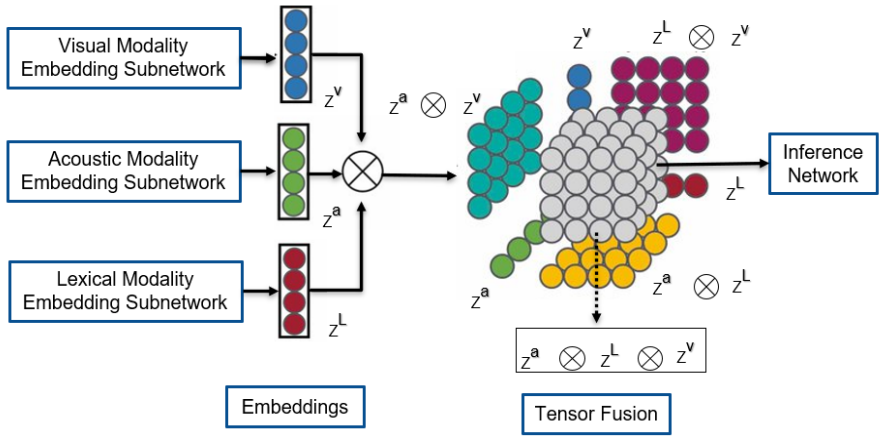


Figure 5: Tensor Fusion Network

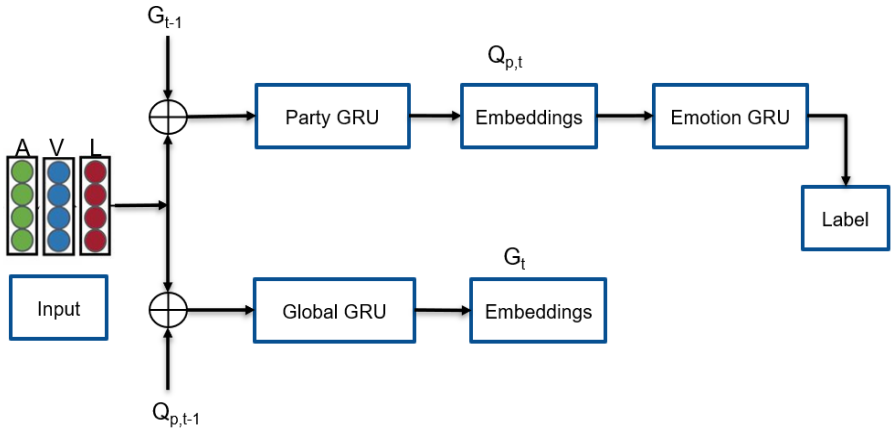


Figure 6: Dialogue RNN

the conversational context (G_t representing global context state at time t) summarized by the global GRU through an attention mechanism. Finally, the updated speaker state is fed into the emotion GRU which models the emotional information for classification.

5.4 Experimental Setup

ADAM optimizer [20] and a class balanced cross-entropy loss [17] were employed during the training for handling dataset imbalance. The number of epochs experimented with during training was in the range of 20 to 300. The learning rate of the optimizer was varied from 1e-05 to 1e-03.

6 BASELINE MODELS

Several baselines were considered to compare our results. First, a text-CNN network [10] which applies CNN to the input utterances without considering the context of the conversation. Then, bcLSTM [10] which represents context using a bi-directional RNN and a Dialogue RNN from the study [10] with effective mechanisms to model context by tracking individual speaker states throughout the conversation for emotion classification. For comparing our multimodal networks, a reinforcement learning method named EDRLF from study [21] and a ConGCN network that models the context-sensitive dependence and speaker-sensitive dependence with graph convolutional network [22].

7 RESULTS

The problem of emotion detection can be thought similar to a multi-class classification problem. Hence, evaluation metrics like weighted average F1 score and class-wise F1 scores are used to measure the performance of our models. A comparison between the Uni-modal models is showcased in Table 2 and for Bi-modal models in Table 3. The results for the multi-modal baselines considering five emotion class labels are described in Table 4 and those considering seven emotion class labels are presented in Table 5.

Models	Modality	Anger	Joy	Surprise	Sadness	Neutral	Weighted F1
MLP	Acoustic	**0.49**	0.16	**0.27**	**0.30**	0.47	0.423
Dialogue RNN [10]	Acoustic	0.34	**0.18**	0.16	0.16	**0.66**	**0.44**
MLP	Lexical	**0.44**	0.52	0.36	**0.61**	0.57	0.50
Dialogue RNN [10]	Lexical	0.41	**0.53**	**0.47**	0.21	**0.77**	**0.60**
MLP	Visual	0.39	0.39	0.21	0.31	0.25	0.31

Table 2: Uni-modal Networks test-set results: class wise & weighted F1-scores.

Models	Modality	Anger	Joy	Surprise	Sadness	Neutral	Weighted F1
MLP + Early Fusion	Acoustic + Lexical	0.50	0.37	0.29	0.47	0.49	0.46
MLP + Late Fusion	Acoustic + Lexical	**0.51**	0.52	0.38	**0.52**	0.61	0.52
Dialogue RNN [10]	Acoustic + Lexical	0.48	**0.53**	**0.48**	0.20	**0.77**	**0.61**
MLP + Early Fusion	Visual + Lexical	**0.32**	**0.44**	**0.19**	0.32	**0.27**	**0.31**
MLP + Late Fusion	Visual + Lexical	0.22	0.10	0.18	**0.41**	0.14	0.20
MLP + Early Fusion	Visual + Acoustic	0.31	**0.46**	**0.19**	**0.34**	0.23	**0.31**
MLP + Late Fusion	Visual + Acoustic	**0.35**	0.19	0.15	0.17	**0.25**	0.23

Table 3: Bi-modal Networks test-set results: class wise & weighted F1-scores.

Models	Anger	Joy	Surprise	Sadness	Neutral	Weighted F1
MLP + Early Fusion	0.51	0.51	0.37	0.48	0.52	0.49
MLP + Late Fusion	0.51	0.54	0.58	0.41	0.61	0.54
CNN-LSTM + Early Fusion	0.52	0.53	0.40	0.47	0.51	0.51
CNN-LSTM + Late Fusion	0.53	0.56	**0.59**	0.40	0.62	0.55
TFN[1]	0.44	0.52	0.36	**0.61**	0.57	0.50
Dialogue RNN[1]	**0.55**	**0.57**	0.43	0.45	**0.75**	**0.62**

[1] represents our implementation of a variant of the original model.

Table 4: Multi-modal Networks test-set results: class wise & weighted F1-scores considering five emotion labels.

Models	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Weighted F1
EDRLF (max) [21]	**0.48**	0	0	**0.55**	**0.77**	0.31	**0.53**	**0.60**
ConGCN [22]	0.47	**0.11**	**0.9**	**0.53**	**0.77**	0.29	0.50	0.59
Dialogue RNN[2]	0.45	0	0.03	0.52	0.76	**0.39**	0.47	0.595

[2] represents our implementation of a variant of the original model.

Table 5: Multi-modal Networks test-set results: class wise & weighted F1-scores considering seven emotion labels.

8 OBSERVATIONS

8.1 Feature Embeddings

During our experiments, we compared the performance of different feature embeddings. We noticed that BERT embeddings achieved a higher F1 score as compared to GLOVE embeddings. The reason behind it could be due to capturing of both left and right context. We also found that utilizing only the functionals such as means, centroid, peaks, etc, instead of the dense representation improved the F1 score significantly. A slight difference was observed between the results after using LSTM encoders compared to the mean pooling approach for visual features.

8.2 Experimental Setup

The combination of the Down sampling technique with a class-balanced loss achieves a higher F1 score when compared to the Up sampling technique. Another interesting thing we observed, the Batch Normalization technique improved the multi-modal network's results but didn't have much effect on uni-modal networks. We also found that down sampling the training samples for the Neutral class helped in avoiding misclassification of other class labels to Neutral. This also resulted in a similar F1 score for all class labels.

8.3 Uni-modal Networks

Considering uni-modal networks belonging to the acoustic modality, we observed that our MLP network outperforms the state-of-the-art Dialogue RNN method [10] for emotions: Anger, Surprise and Sadness and gives a close weighted F1 score due to improved features and technique of class balancing and class balanced loss. For the uni-modal networks belonging to the lexical modality, the state-of-the-art Dialogue RNN model [10] achieves a very high f1 score for the Neutral emotion class due to a large number of training samples, whereas our model achieves similar results for all class labels. For Visual modality uni-modal networks, we observed that the Neutral class has a low f1 score which, could be due to ambiguity in the cues extracted for the training samples belonging to this class.

8.4 Bi-modal Networks

Considering the combination of acoustic and lexical modalities, we observed that the state-of-the-art Dialogue RNN method [10] outperforms our Late fusion Model. We noticed another interesting observation that our model outperforms for sadness class label even though it had the least number of training samples. For models involving the visual modality, early fusion seems to work better than late fusion.

8.5 Multi-modal Networks

Our implementation of a variant of the Dialogue RNN model [6] outperforms all the other models considering five emotion classes from the dataset. We also compared our best model's results with other state-of-the-art methods for all seven emotion classes from the dataset and found that the EDRLF (max) method from the study [21] achieves the highest weighted F1 score.

9 LESSONS LEARNT

We learned a variety of lessons throughout this project. The following are the key concepts we grasped while studying and experimenting with different state-of-the-art models. We understood that additive fusion techniques like early fusion and late fusion rely on figuring out relative emphasis on various modalities. However, in reality, we cannot rely on every modality due to noises like sensor noise and many others. We also understood how Tensor Fusion explicitly models uni-modal, bi-modal, and multi-modal interactions and their impact on our problem. Lastly, we found that we require the context, state of the speaker, state of the previous speaker, and emotion of the previous speaker to determine the emotions of the speaker of the utterance. Dialogue RNN captures all these aspects efficiently by maintaining individual party states and global context states throughout the conversation.

10 CONCLUSION

Lexical uni-modal networks achieve the highest F1 score as they capture the context in a conversation for emotion recognition. But these conversations could be ambiguous like we saw in Fig 4. To handle this, we experimented with various bi-modal networks and found that acoustic features when combined with lexical perform best. We also observed cases with contradictory cues in these two modalities, so we shifted to multi-modal networks. Finally, we concluded that multi-modal interactions capture facial expressions, tone, context, and other cues to determine the emotion label efficiently.

11 CONTRIBUTION

- **Mehak:**
  Statistical analysis, literature review, acoustic feature extraction, creating a subset of data with a balanced number of training samples for each class label, implementing and fine-tuning various models like uni-modal networks for the acoustic modality, bi-modal networks (early and late fusion) for acoustic and lexical pair, early fusion and late fusion models for CNN-LSTM multi-modal networks, and the Tensor Fusion Network, designing the final presentation, creating model diagrams, designing the final report.

- **Disha:**
  Lexical feature extraction, creating a subset of data with a balanced number of samples for each class label, implementing and fine-tuning various models like uni-modal networks for the lexical modality, bi-modal networks (early and late fusion) for lexical and visual pair, and early fusion and late fusion models for MLP multi-modal networks, and designing the mid-term presentation.

- **Rhushabh:**
  Visual feature extraction, creating a subset of data with a balanced number of samples for each class label, implementing and fine-tuning various models like uni-modal networks for the visual modality, bi-modal networks (early and late fusion) for acoustic and visual pair, and Dialogue RNN multi-modal network (for both five emotion class labels and seven emotion class labels).

REFERENCES

[1] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *CoRR*, abs/1905.02947, 2019.

[2] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: commonsense knowledge for emotion identification in conversations. *CoRR*, abs/2010.02795, 2020.

[3] M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. 34.

[4] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, Jul. 2019.

[7] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing (EMNLP-IJCNLP), pages 154–164, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Saurav Sahay, Shachi H Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. Multimodal relational tensor network for sentiment and emotion classification. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 20–27, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] Aman Shenoy and Ashish Sardana. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA, July 2020. Association for Computational Linguistics.

[10] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. pages 527–536, 01 2019.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery.

[12] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Representation learning for speech emotion recognition. pages 3603–3607, 09 2016.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[16] Qi Fu, Shiwei Ma, Lina Liu, and Jinjin Liu. Human action recognition based on sparse lstm auto-encoder and improved 3d cnn. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 197–201, 2018.

[17] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *CoRR*, abs/1901.05555, 2019.

[18] L. Wikarsa and S. N. Thahir. A text mining application of emotion classifications of twitter's users using naïve bayes method. In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pages 1–6, 2015.

[19] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Xiangdong Huang, Minjie Ren, Qiankun Han, Xiaoqi Shi, Jie Nie, Weizhi Nie, and An-An Liu. Emotion detection for conversations based on reinforcement learning framework. *IEEE MultiMedia*, pages 1–1, 2021.

[22] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. pages 5415–5421, 08 2019.