Multimodal Emotion Recognition in Conversations: a Literature Review

Deep Learning on Public Dataset
CM3070, Final Computer Science Project

Hudson Leonardo MENDES
hlm12@student.london.ac.uk
University of London
2023

Page Count (excluding References): 6 pages

Abstract

Emotion Recognition in Conversations (or "ERC") is a complex and relatively unaddressed problem, with the best solution only reaching a highest F1-score of 67.25% for the reference Benchmark Dataset, MELD. The MELD dataset is investigated alongside some of the most representative models in the sparse ERC solution space, namely the DialogRNN, M2FNet and the SPCL-CL-ERC, in their commonalities, architectures, results and insights, and how they approach addressing emotion classification on multi-party non-dyadic settings. A critique attempts to correlate and contrast how each one of the solutions proposed address the ERC problem in different ways, and attempts to highlight intuitive gaps that could be explored by follow-up research.

Table of Contents

Introduction

Emotion Recognition in Conversations (or "**ERC**") recurringly appears classified by researchers in the field as a relatively complex task, especially in non-dyadic settings where there are more than two people engaging in a Dialog.

Poria et al (2019) describe that the ERC task "presents several challenges such as conversational context modelling, emotion shift of the interlocutors"[1] and that these challenges "make the task more difficult to address"[1]. They illustrate a situation stating that "[u]tterances like 'yeah', 'okay', 'no' can express varied emotions depending on the context and discourse of the dialogue (…) most models resort to assigning the majority class"[1].

However, despite its challenges, the vast applications of ERC[2] including "opinion mining over chat history and social media threads in YouTube, Facebook, Twitter"[3] (Majumder et al, 2019), continues to draw attention to its problem space, given that any improvements in the field can provide the industries that they serve with significant breakthrough.

The present literature review explores papers of solutions proposed for the ERC problem space, specifically the ones trained and evaluated over the MELD Dataset[1], breaks down the models proposed into their building blocks, and draws a comparison of components that are present in each one of them, and highlights possible implementation gaps that can be explored for follow-up experimentation.

Scope of Research & Sparsity of ERC Solution Space

Many more models proposed to solve the problem of Emotion Recognition in Conversation than the present literature review could possibly analyse, and a criteria to prioritise the models, and their respective literature, has been devised focusing exclusively on solutions introduced by papers which have reported against the ECR over MELD benchmark[10].

To date, 45 papers have addressed ERC and published results against the MELD dataset, not a relatively large number of papers when compared to, for example, text classification over the GLUE dataset (301 papers), the

complexity of multi-modal models is often non-trivial and each paper requires significant effort to be investigated.

These 45 papers describe models clustering together as being either multi-modal models and textual-only models. And although they have been devised to solve the same problem (ERC), their architecture varies heavily creating a vast yet sparsely explored solution space, with many possible combinations of their building blockers yet to be explored.

The present literature review focuses on models that are reference to the MELD dataset, such as the DialogRNN[3] used as baseline to the MELD dataset paper itself, and to models that are both the highest ranking and that are somehow a prototypical representation to other models in their clusters, such as the top-ranking text-only SPCL-CL-ERC[8], and the M2FNet[6] which is the highest ranking multi-modal model in the ranking[10]. Underlying research required to make sense of these three models were also investigated and incorporated the conclusions hereby documented.

Definition of "Emotions" in the context of ERC

"Emotions are the unseen mental states that are linked to thoughts and feelings"[4]. And while there have been some breakthroughs in using brain activity for inference, these could implicate serious ethical concerns and still require significant equipment to be attached to the person[5], not viable to the applications for which ERC is relevant[2].

From that limitation, Chudasama et al. (2022) concludes that "[i]n the absence of physiological indications, they could only be detected by human actions such as textual utterances, visual gestures, and acoustic signals"[6], which indicates that to solve ERC effectively, one must must take in consideration the multi-modality of ERC data.

Multiple authors seem to agree that exploring multi-modality is important to solve the ERC problem[6][3][7]. However many others utilise a single modality, usually textual[8][9], and even without leveraging data from the other modalities, still outperform multimodal models, as for example the SPCL-CL-ERC[8], the current state-of-the-art in ERC over the MELD dataset.

Research Paper 1: MELD Dataset

EmotionLines is an emotion corpus of multi-party (non-dyadic) conversations developed by Chen et al. (2018). "The MELD dataset has evolved from the EmotionLines dataset"[1] and delivered a significant number of improvements. It's based on dialogs extracted from the TV Series "Friends" containing two or more people engaging in dialogue.

MELD, opposed to EmotionLines, "provides multimodal sources"[1] and over 13,000 utterances. It was re-annotated using "Ekman's six universal emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust)"[1] and "two additional emotion labels: Neutral and Non-Neutral"[1].

The authors of MELD also attempted to improve the annotation process. For EmotionLines, was based on Amazon Mechanical Turk (AMT) where annotators only looked at the transcriptions and their domain over the English language could not be asserted. For MELD, "[t]he annotators were graduate students with high proficiency in English speaking and writing"[1] and "were briefed about the annotation process with a few examples"[1]. The improved process led MELD to reach a Fleiss kappa score of 0.43 against 0.34 achieved by EmotionLines.

Consequently, the MELD dataset has become one of the most important benchmark datasets for Emotion Recognition in Conversations (ERC) with 45 papers publishing metrics against its test data split[10].

Research Paper 2: DialogRNN Model

The first model investigated is the DialogRNN used as the baseline model in the MELD Dataset paper[1]. This model presents a series of GRUs (or "Gated Recurrent Units") that update different states, individually, and all feed into one another to perform the emotion classification task[3].

The entire model can be formalised by the following set of equations equation:

- Let the t be a given timestep, u_t be the utterance the timestep t
- Let $T(u_t)$, $V(u_t)$, $A(u_t)$ be the functions that performs textual feature extraction, video feature extraction and audio feature extraction respectively on a given utterance[3];

- X_t be the concatenation of the text, audio and video representations at time t into a single feature vector[3];
- Let $q_{P,t}$ be the state of any individual speaker P, g_{t-1} be the dialogue context state, and e_t be the emotion label state[3];
- Let $GRU_P(g_{t-1}, X_t)$, $GRU_G(g_{P,t-1}, X_t)$ and $GRU_E(e_{t-1}, q_{P,t})$ be the Gated Recurrent Units responsible for generating representations for the individual speaker state, to the global dialogue state and to the individual speaker emotion states respectively[3];

The prediction y_{pred} of the emotion state e_t is given by the following equation:

$$X_t = T(u_t) + V(u_t) + A(u_t)$$
$$q_{P,t} = GRU_P(g_{t-1}, X_t)$$
$$g_{t-1} = GRU_G(g_{P,t-1}, X_t)$$
$$y_{pred} = e_t = GRU_E(e_{t-1}, q_{P,t})$$[3]

As it's possible to observe from the set of equations above, the y_{pred} is mapped from a number of chained dependencies. Namely the g_{t-1}, which is the previous dialogue global state used to compute the $q_{P,t}$, which is the current state of the speaker P, which is then finally input into the GRU_E, alongside the previous emotion states to generate the emotion label $y_{pred} = e_t$.

The authors also propose and comparatively evaluate variations of their base model architecture. The variant $BiDialogueRNN + at_{MM}$ is the best performing one, and makes use of an extra Attention layer responsible for learning relationships between individual utterances[3].

During the ablation studies, Majumder et al demonstrates that "party state stands very important, as without its presence the performance falls by 4.33%", whereas the "Emotion GRU['s] (...) absence causes performance to fall by only 2.51%[3], which indicates that somehow keeping record of the individual state of individuals might imply more accurate emotion classification.

Paper 3: M2FNet Model

Chudasama et al introduced in 2022 a multi-modal fusion network model (or "M2FNet") which "takes advantage of the multi-modal nature of real-world media content" and by combining "features from different modalities to generate rich emotion-relevant representations"[6].

This model largely leverages the power of the attention mechanism in multiple levels of its architecture, including (a) the Text Encoding which is based on a feature map generated by a modified version of the RoBERTa encoder, the $\phi_M - RoBERTa$[6], as well as (b) a "multi-head attention-based fusion layer is introduced, which aids the proposed system to combine latent representations of the different inputs"[6].

While DialogRNN does not utilise pre-trained text encoders, M2FNet leverages the pretrained features provided by $RoBERTa$ introduced by Zhuang et al. in 2021 is a pre-trained encoder which is leveraged to produce text feature maps, while DialogRNN

To explore the audio and video modalities, the M2FNet also introduces two non-trivial feature extraction models, which also leverage attention architectures to produce feature maps, for Video Feature Extraction and Audio Feature Extraction, both trained with a newly proposed "adaptive margin-based triplet loss function"[6].

The feature-maps from all modalities are then fed into the fusion-network that produces a representation which is then used for the emotion classification.

Differently to the DialogRNN model, the M2FNet does not model the individual speaker context separately to the dialogue context. Instead it suggests a (a) "utterance level feature extraction" stage, where the multi-modal features for each one the utterances is extracted separately, followed by a (b) "dialogue level feature extraction", where "Each modality embeddings (...) are passed through their corresponding network with a variable stack of transformer encoders 32 to learn the inter utterance context"[6]. Both the "utterance level" (through a residual connection[12]) and "dialogue level" features are then input into the fusion mechanism.

M2FNet is a large and complex model and, due to its complexity, it is not possible to capture enough of its features in a simplified set of equations that describe it elegantly without missing on important details of its

architecture. Given its vast use of transformer models in its ensemble, M2FNet may also be computationally considerably expensive, but its computational complexity is neither provided in the paper nor by related work.

Chudasama et al claims that results demonstrate that their "fusion mechanism helps to enhance the accuracy (...) for IEMOCAP and MELD datasets" and "$F1_{score}$ for the IEMOCAP dataset"[6], and the ablation tests indicate that 5 "Attention Fusion Layers" slightly outperformed 6 "Attention Fusion Layers".

Additionally, the paper proposes that features extracted from both the scene and faces can be combined using a "weighted Face Model", which outperformed representations created by either feature source on its own.

Paper 4: SPCL-CL-RC Model

In opposition to DialogRNN and M2FNet previously described, the novelty introduced by the SPCL-CL-RC paper (Song et al, 2022) is not a substantially different model architecture, but instead a loss function, the "Supervised Prototypical Contrastive Learning (SPCL) loss"[8] which "ensures that each sample has at least one positive sample of the same category and negative samples of all other categories within a mini-batch"[8] combined with "curriculum learning (...) with contrastive learning"[8].

Curriculum Learning is a training approach which states that hypothetically "a well chosen curriculum strategy can act as a continuation method (...), i.e., can help to find better local minima of a non-convex training criterion"[13]. Song et al achieve that by designing a "sorting the training data via this function, we can schedule the training data in an easy-to- hard fashion"[8].

Similarly to M2FNet, and differently to DialogRNN, SPCL-CL-ERC uses pre-trained encoder SimCSE model introduced by Gao et al. in 2021[9] and leverages the Transformer architecture, but that also trains sentence representation using a Contrastive Learning Approach[9] to improve the separation between embeddings of contradictory utterances.

And even though SPCL-CL-ERC does not leverage the multi-modality of the MELD dataset, it ranks first in the With Code Ranking[10], with an accuracy 0.81% better than M2FNet which does leverage the multi-modality of the dataset.

Worth of mention, SPCL-CL-ERC also makes use of a technique recently popularised, novel in the context of ERC, called prompt engineering[15] which is a text generation technique based on the “pre-train, prompt, and predict” approach[15] and has been employed to introduce individual speaker context to the dialogue context using a "prompt-based context encoder"[8].

The algorithm proposed can be summarised in the following way:

- Let ε be the label set of emotions;
- Let $P_t = for\ u_t, s_t\ fells < mask >$ be the prompt composed, u_t be the textual utterance, all at time t[8];
- Let $C_t = [(s_{t-k}, u_{t-k}), (s_{t-k+1}, u_{t-k+1}), \ldots, (s_t, u_t)]$ be the dialogue context at time t composed by k tuples (s_t, u_t)[8];
- Let $H_t^k = SimCSE(C_t \oplus P_t)$ be the model that takes in a sequence of prompts and returns hidden states for each token at time t, the H_t^k[8];
- Let $G(z_i, z_j)$ be the cosine similarity and τ be a scalar temperature parameter[8]
- Let $I = [z_1, z_2, \cdots, z_N]$ be the batch of size N with all the representations produced by the context encoder, extracted from H_t^k, $A(i) = \{z_x \in I \mid x \neq i\}$ the set of all encoded representations different to z_i, and $P(i) = \{z_x \in I \mid y_x = y_i\}$ the set of all positive examples[8], $E(i) = |\{\varepsilon_i \in \varepsilon \mid \varepsilon_i = y_i\}|$ the set of all emotion labels equal to y_i;
- Let $Q_i = [z_1^i, z_2^i, \ldots, z_M^i]$ be the representation queue containing at most the M most recent representations generated by the $SimCSE(C_t \oplus P_t)$ encoder;

The total loss L_{spcl} is computed with modified versions of the the contrastive loss function components $N_{sup}(i)$ and $P_{sup}(i)$, named the $N_{spcl}(i)$ and $P_{spcl}(i)$ respectively:

$$F(z_i, z_j) = exp(G(z_i, z_j)/\tau)$$

$$N_{sup}(i) = \sum_{z_j \in A(i)} F(z_i, z_j)$$

$$\rho_{sup}(i) = \sum_{z_j \in P(i)} F(z_i, z_j)$$

$$S_k = RandomSelect(Q_i, k)$$

$$T_k = \frac{1}{k} \sum_{z_j^i \in S_k, j \in [1..k]}^{i} z_j^i$$

$$Ty_i = \frac{1}{k} \sum_{z_i \in P(i)}^{i} z_i$$ [see footnote #1]

$$N_{spcl}(i) = N_{sup}(i) + \sum_{k \in E(i)}^{i} F(z_i, T_k)$$

$$P_{spcl}(i) = P_{sup}(i) + F(z_i, Ty_i)$$

$$L_i^{spcl} = -log(\frac{1}{P(i)} \frac{P_{spcl}(i)}{N_{spcl}(i)})$$

$$L_{spcl} = \sum_{i=0}^{N} L_i^{spcl}$$ [8]

To predict the emotion label SPCL authors "train an additional linear layer to predict the labels using cross-entropy loss"[8] that utilise the fine-tuned SimCSE representations for classification.

Given that the SPCL-CL-ERC model does not operate on all modalities, it also relies on curriculum learning to eliminate extreme cases (most ambiguous) from training examples of its encoder and, to achieve that, employs a "difficulty measure"[8] that determines that "the closer the sample is to the category centre, the lower the difficulty"[8]. The authors provide a qualitative analysis where they claim that a sample-based curriculum learning strategy has proven to be more efficient than direct sorting of the data using the difficulty function which led the "model [to] overfit on simple samples in the early stage of training and produce large losses in the later stage"[8].

The paper states that results demonstrate that the "proposed SPCL and curriculum learning strategy, (...) achieve[d] state-of-the-art results on three benchmarks"[8].

Critique

All papers investigated base their ideas on solid literature, all of them providing comparative results between theirs and existing alternative solutions, as well as open sourced their code. They also provide reasonable ablation studies demonstrating how each one of the underlying components brought into their models provide the benefits they were intended for.

RNNs (and similarly GRU and LSTM), present in the architecture of the DialogRNN model, are known to be inefficient in learning longer-range dependencies in due to the fact that "gradient descent becomes increasingly inefficient when the temporal span of the dependencies increases."[16]. However, the authors of DialogRNN do attempt to compensate for that deficiency with an additional Attention mechanism responsible for learning long range dependencies[3].

For natural language processing, RNNs have been largely replaced by transformer-based models, as it is noticeable in the two other models, M2FNet and SPCL-CL-ERC, introduced more recently at a time that the benefits of attention mechanism had over learning long-term dependencies in text[6][8].

While M2FNet deploys a hierarchy of several transformer models with multiple attention heads[6], SPCL-CL-ERC demonstrates that state-of-the-art accuracy can be reached operating only over the textual modality alone[8][10], with a single transformers based model (SimCSE) if optimised appropriately, which could indicate there is unnecessary complexity in the M2FNet ensemble that could be eliminated to produce a simpler and more efficient model. Results on different benchmarks demonstrate that M2FNet performs better than SPCL-CL-ERC for different datasets[18].

Another noteworthy aspect is that modelling a dialogue so that it is observed as a sequence of interactions, and that the arising emotions are a function of our internal

[1] The equation for the term Ty_i has not been provided by the author, and it was inferred based on the information provided in the paper[8] as well in the source code provided with the paper, source: https://github.com/caskcsg/SPCL/blob/cd489397df93dc43bbe20e0e57b0814438666ad9/spcl_loss.py

states produced by that sequence seems to draw from a shared robust intuition. DialogRNN encodes that intuition directly into its architecture by making use of GRUs[3] that always utilise previous states when predicting new states. M2FNet also encodes that in its architecture by passing echo modality utterance embedding "through their corresponding network with a variable stack of transformer encoders (...) to learn the inter utterance context"[6].

Conversely, SPCL-CL-ERC has a fairly unique approach to modelling sequences of utterances as previous context to any given utterance using prototypical representations which does not attempt to learn the relationships between the the present and previous utterances, but that does encode the sequential dialog intuition into representations[8].

Whereas it's debatable whether the sequence of interactions, individual relationship of utterances, and emotions as functions of that sequence, are or aren't fundamental aspects of dialogues, different data distributions with different densities for any of these features may show that one model architecture responds to change and is more robust than the others.

A common issue across them all is that none of the papers discusses the distributions of data available in the benchmark datasets and whether they are representative of multi-party conversation scenarios different to the ones captured in the TV Series "Friends". Different data skews could significantly limit any model's ability to operate on a different dataset.

Given SPCL-CL-ERC's application of Curriculum Learning, which eliminates extreme cases[8], and the unique way in which it models sequences of previous utterances through prototypical representations, different distributions in alternative ERC datasets could render SPCL-CL-ERC particularly vulnerable. Any given datasets that present a higher density of ambiguous cases, or that presents dialogs that rely more heavily in previous context to determine a present emotion, might show that the rather simplistic approach proposed by SPCL-CL-ERC may not be as efficient as its multi-modal counterparts capable of leveraging features of other modalities for disambiguation or of learning relationships between individual utterances, like the DialogRNN and M2FNet, a situation that occurred in at least one situation: for the IEMOCAP dataset[18].

Conclusion

The challenging problem of Emotion Recognition in Conversation has yet much room for exploration, with the best performing model (SPCL-CL-ERC) at 67.25% F1-score.

Architectures proposed to model ERC vary drastically[3][6][8]. Many designs were attempted, ranging from GRUs[3] to ensembles of multiple transformer models stacked[8]. Still some of the highest metrics were achieved by models that do not leverage multi-modality[8] and do not attempt to encode a fundamental intuition understanding dialogs requires taking in consideration relationships between multiple utterances in a dialog[8].

Prompt-based context modelling was just tangentially used across the available models[8], possibly due to its recency but can already be found in the ERC literature.

Finally, the MELD dataset with its improvements against the EmotionLines on top of which it builds delivers a good starting point for researchers committed to drive improvements to the ERC forwards.

References

[1] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 527–536. DOI:https://doi.org/10.18653/v1/P19-1050

[2] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems 115, (2018), 24–35. DOI:https://doi.org/https://doi.org/10.1016/j.dss.2018.09.002

[3] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence 33, 1 (2019), 6818–6825. DOI:https://doi.org/10.1609/aaai.v33i01.33016818

[4] Soujanya Poria, Navonil Majumderd, Rada Mihalceae, and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. IEEE Access PP, (July 2019), 1–1. DOI:https://doi.org/10.1109/ACCESS.2019.2929050

[5] Wenyi Li, Shengjie Zheng, Yufan Liao, Rongqi Hong, Chenggang He, Weiliang Chen, Chunshan Deng, and Xiaojian Li. 2023. The brain-inspired decoder for natural visual image reconstruction. Frontiers in Neuroscience 17, (2023). DOI:https://doi.org/10.3389/fnins.2023.1130606

[6] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4651–4660. DOI:https://doi.org/10.1109/CVPRW56347.2022.00511

[7] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 5415–5421. DOI:https://doi.org/10.24963/ijcai.2019/752

[8] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5197–5206. Retrieved from https://aclanthology.org/2022.emnlp-main.347

[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings (EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings), Association for Computational Linguistics (ACL), 6894–6910.

[10] Papers with Code, "Emotion Recognition in Conversation on MELD". Source: https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld

[11] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. arXiv preprint arXiv:2108.12009, 2021.

[12] Francois Chollet. 2017. Deep Learning with Python (1st ed.), "Going beyond the Sequential model: the Keras functional API", pp. 244. Manning Publications Co., USA.

[13] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/1553374.1553380

[14] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 1218–1227. Retrieved from https://aclanthology.org/2021.ccl-1.108

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv. 55, 9, Article 195 (September 2023), 35 pages. https://doi.org/10.1145/3560815

[16] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5, 2 (1994), 157–166. DOI:https://doi.org/10.1109/72.279181

[17] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan. Retrieved from https://aclanthology.org/L18-1252

[18] Papers with Code, "Emotion Recognition in Conversation on IEMOCAP". Source: https://paperswithcode.com/sota/emotion-recognition-in-conversation-on