MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations

Soujanya Poria[†], **Devamanyu Hazarika**[Φ], **Navonil Majumder**[‡], **Gautam Naik**[¶], **Erik Cambria**[¶], **Rada Mihalcea**[ι]
[†]Information Systems Technology and Design, SUTD, Singapore
[Φ]School of Computing, National University of Singapore, Singapore
[‡]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[¶]Computer Science & Engineering, Nanyang Technological University, Singapore
[ι]Computer Science & Engineering, University of Michigan, USA
sporia@sutd.edu.sg, hazarika@comp.nus.edu.sg,
navo@nlp.cic.ipn.mx, gautam@sentic.net,
cambria@ntu.edu.sg, mihalcea@umich.edu

Abstract

Emotion recognition in conversations (ERC) is a challenging task that has recently gained popularity due to its potential applications. Until now, however, there has been no large-scale multimodal multi-party emotional conversational database containing more than two speakers per dialogue. To address this gap, we propose the *Multimodal EmotionLines Dataset* (MELD), an extension and enhancement of EmotionLines. MELD contains about 13,000 utterances from 1,433 dialogues from the TV-series *Friends*. Each utterance is annotated with emotion and sentiment labels, and encompasses audio, visual, and textual modalities. We propose several strong multimodal baselines and show the importance of contextual and multimodal information for emotion recognition in conversations. The full dataset is available for use at http://affective-meld.github.io.

1 Introduction

With the rapid growth of Artificial Intelligence (AI), multimodal emotion recognition has become a major research topic, primarily due to its potential applications in many challenging tasks, such as dialogue generation, user behavior understanding, multimodal interaction, and others. A conversational emotion recognition system can be used to generate appropriate responses by analyzing user emotions (Zhou et al., 2017; Rashkin et al., 2018).

Although significant research work has been carried out on multimodal emotion recognition using audio, visual, and text modalities (Zadeh et al., 2016a; Wollmer et al., 2013), significantly less work has been devoted to emotion recognition in conversations (ERC). One main reason for this is the lack of a large multimodal conversational dataset.

According to Poria et al. (2019), ERC presents several challenges such as conversational context modeling, emotion shift of the interlocutors, and others, which make the task more difficult to address. Recent work proposes solutions based on multimodal memory networks (Hazarika et al., 2018). However, they are mostly limited to dyadic conversations, and thus not scalable to ERC with multiple interlocutors. This calls for a multi-party conversational data resource that can encourage research in this direction.

In a conversation, the participants' utterances generally depend on their conversational context. This is also true for their associated emotions. In other words, the context acts as a set of parameters that may influence a person to speak an utterance while expressing a certain emotion. Modeling this context can be done in different ways, e.g., by using recurrent neural networks (RNNs) and memory networks (Hazarika et al., 2018; Poria et al., 2017; Serban et al., 2017). Figure 1 shows an example where the speakers change their emotions (emotion shifts) as the dialogue develops. The emotional dynamics here depend on both the previous utterances and their associated emotions. For example, the emotion shift in utterance eight (in the figure) is hard to determine unless cues are taken from the facial expressions and the conversational history of both speakers. Modeling such complex inter-speaker dependencies is one of the major challenges in conversational modeling.

Conversation in its natural form is multimodal. In dialogues, we rely on others' facial expressions, vocal tonality, language, and gestures to anticipate their stance. For emotion recognition, multimodal-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536
Florence, Italy, July 28 - August 2, 2019.

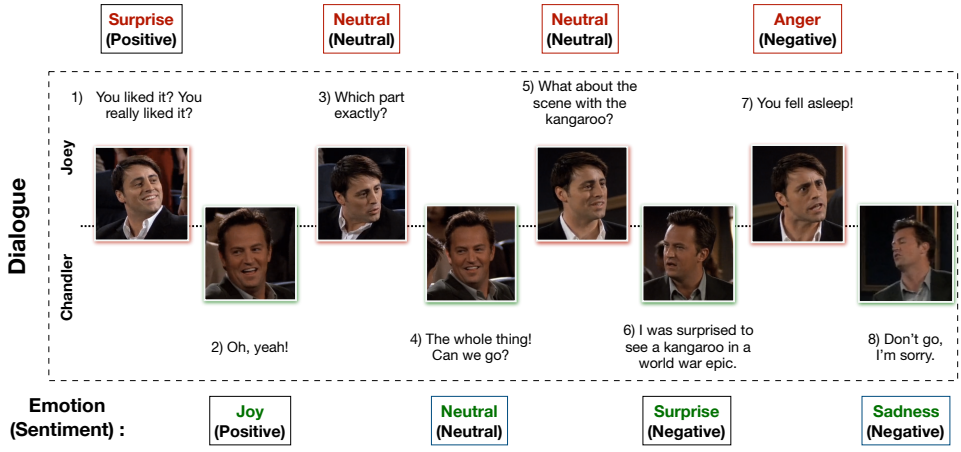

Figure 1: Emotion shift of speakers in a dialogue in comparison with their previous emotions.

Utterance: *"Become a drama critic!"*
Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*
Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

Figure 2: Importance of multimodal cues. Green shows primary modalities responsible for sentiment and emotion.

ity is particularly important. For the utterances with language that is difficult to understand, we often resort to other modalities, such as prosodic and visual cues, to identify their emotions. Figure 2 presents examples from the dataset where the presence of multimodal signals in addition to the text itself is necessary in order to make correct predictions of their emotions and sentiments.

Multimodal emotion recognition of sequential turns encounters several other challenges. One such example is the classification of short utterances. Utterances like "*yeah*", "*okay*", "*no*" can express varied emotions depending on the context and discourse of the dialogue. However, due to the difficulty of perceiving emotions from text alone, most models resort to assigning the majority class (e.g., *non-neutral* in EmotionLines). Approximately 42% of the utterances in MELD are shorter than five words. We thus provide access to the multimodal data sources for each dialogue and posit that this additional information would benefit the emotion recognition task by improving the context representation and supplementing the missing or misleading signals from other modalities. Surplus information from attributes such as the speaker's facial expressions or intonation in speech could guide models for better classification. We also provide evidence for these claims through our experiments.

The development of conversational AI thus depends on the use of both contextual and multimodal information. The publicly available datasets for multimodal emotion recognition in conversations – IEMOCAP and SEMAINE – have facilitated a significant number of research projects, but also have limitations due to their relatively small number of total utterances and the lack of multi-party conversations. There are also other multimodal emotion and sentiment analysis datasets, such as MOSEI (Zadeh et al., 2018), MOSI (Zadeh et al., 2016b), and MOUD (Pérez-Rosas et al., 2013), but they contain individual narratives instead of dialogues. On the other hand, EmotionLines (Chen et al., 2018) is a dataset that contains dialogues from the popular TV-series *Friends* with more than two speakers. However, EmotionLines can only be used for textual analysis as it does not provide data from other modalities.

In this work, we extend, improve, and further develop the EmotionLines dataset for the multimodal scenario. We propose the *Multimodal EmotionLines Dataset* (MELD), which includes not only textual dialogues, but also their corresponding visual and audio counterparts. This paper makes several contributions:

- MELD contains multi-party conversations that are more challenging to classify than dyadic variants available in previous datasets.
- There are more than 13,000 utterances in MELD,

which makes our dataset nearly double the size of existing multimodal conversational datasets.

- MELD provides multimodal sources and can be used in a multimodal affective dialogue system for enhanced grounded learning.
- We establish a strong baseline, proposed by Majumder et al. (2019), which is capable of emotion recognition in multi-party dialogues by inter-party dependency modeling.

The remainder of the paper is organized as follows: Section 2 illustrates the EmotionLines dataset; we then present MELD in Section 3; strong baselines and experiments are elaborated in Section 4; future directions and applications of MELD are covered in Section 5 and 6, respectively; finally, Section 7 concludes the paper.

2 EmotionLines Dataset

The MELD dataset has evolved from the EmotionLines dataset developed by Chen et al. (2018). EmotionLines contains dialogues from the popular sitcom *Friends*, where each dialogue contains utterances from multiple speakers.

EmotionLines was created by crawling the dialogues from each episode and then grouping them based on the number of utterances in a dialogue into four groups of [5, 9], [10, 14], [15, 19], and [20, 24] utterances respectively. Finally, 250 dialogues were sampled randomly from each of these groups, resulting in the final dataset of 1,000 dialogues.

2.1 Annotation

The utterances in each dialogue were annotated with the most appropriate emotion category. For this purpose, Ekman's six universal emotions (*Joy, Sadness, Fear, Anger, Surprise, and Disgust*) were considered as annotation labels. This annotation list was extended with two additional emotion labels: *Neutral* and *Non-Neutral*.

Each utterance was annotated by five workers from the Amazon Mechanical Turk (AMT) platform. A majority voting scheme was applied to select a final emotion label for each utterance. The overall Fleiss' kappa score of this annotation process was 0.34.

3 Multimodal EmotionLines Dataset (MELD)

We start the construction of the MELD corpus by extracting the starting and ending timestamps of all utterances from every dialogue in the EmotionLines dataset. To accomplish this, we crawl through the subtitles of all the episodes and heuristically extract the respective timestamps. In particular, we enforce the following constraints:

1. Timestamps of the utterances in a dialogue must be in an increasing order.
2. All the utterances in a dialogue have to belong to the same episode and scene.

These constraints revealed a few outliers in EmotionLines where some dialogues span across scenes or episodes. For example, the dialogue in Table 2 contains two natural dialogues from episode 4 and 20 of season 6 and 5, respectively. We decided to filter out these anomalies, thus resulting in a different number of total dialogues in MELD as compared to EmotionLines (see Table 1).

Next, we employ three annotators to label each utterance, followed by a majority voting to decide the final label of the utterances. We drop a few utterances where all three annotations were different, and also remove their corresponding dialogues to maintain coherence. A total of 89 utterances spanning 11 dialogues fell under this category.

Finally, after obtaining the timestamp of each utterance, we extract their corresponding audio-visual clips from the source episode followed by the extraction of audio content from these clips. We format the audio files as 16-bit PCM WAV files for further processing. The final dataset includes visual, audio, and textual modalities for each utterance.[1]

Dataset	# Dialogues			# Utterances		
	train	dev	test	train	dev	test
EmotionLines	720	80	200	10561	1178	2764
MELD	1039	114	280	9989	1109	2610

Table 1: Comparison between the original EmotionLines dataset and MELD.

3.1 Dataset Re-annotation

The utterances in the original EmotionLines dataset were annotated by looking only at the transcripts. However, due to our focus on multimodality, we re-annotate all the utterances by asking the three annotators to also look at the available video clip of the utterances. We then use majority-voting to obtain the final label for each utterance.

[1] We consulted a legal office to verify that the usage and distribution of very short length videos fall under the *fair use* category.

Episode	Utterance	Speaker	Emotion	Sentiment
S6.E4	What are you talkin about? I never left you! Youve always been my agent!	Joey	surprise	negative
	Really?!	Estelle	surprise	positive
	Yeah!	Joey	joy	positive
	Oh well, no harm, no foul.	Estelle	neutral	neutral
S5.E20	Okay, you guys free tonight?	Gary	neutral	neutral
	Yeah!!	Ross	joy	positive
	Tonight? You-you didn't say it was going to be at nighttime.	Chandler	surprise	negative

Table 2: A dialogue in EmotionLines where utterances from two different episodes are present. The first four utterances in this dialogue have been taken from episode 4 of season 6. The last three utterances in red font are from episode 20 of season 5.

The annotators were graduate students with high proficiency in English speaking and writing. Before starting the annotation, they were briefed about the annotation process with a few examples.

We achieve an overall Fleiss' kappa score of 0.43 which is higher than the original EmotionLines annotation whose kappa score was 0.34 (kappa of IEMOCAP annotation process was 0.4), thus suggesting the usefulness of the additional modalities during the annotation process.

2,772 utterances in the EmotionLines dataset were labeled as *non-neutral* where the annotators agreed that the emotion is not neutral but they could not reach agreement regarding the correct emotion label. This hampers classification, as the *non-neutral* utterance space and the other emotion-label spaces get conflated. In our case, we remove the utterances where the annotators fail to reach an agreement on the definite emotion label.

The number of disagreements in our annotation process is 89, which is much lower than the 2,772 disagreements in EmotionLines, reflecting again the annotation improvement obtained through a multimodal dataset. Table 3 shows examples of utterances where the annotators failed to reach consensus.

Table 4 shows the label-wise comparison between EmotionLines and MELD dataset. For most of the utterances in MELD, the annotations match the original annotations in EmotionLines. Yet, there exists a significant amount of samples whose utterances have been changed in the re-annotation process. For example, the utterance *This guy fell asleep!* (see Table 5), was labeled as *non-neutral*

Utterance	Annotator 1	Annotator 2	Annotator 3
You know? Forget it!	sadness	disgust	anger
Oh no-no, give me some specifics.	anger	sadness	neutral
I was surprised to see a kangaroo in a World War epic.	surprise	anger	joy
Or, call an ambulance.	anger	surprise	neutral

Table 3: Some examples of the utterances for which annotators could not reach consensus.

	Categories	EmotionLines Train	EmotionLines Dev	EmotionLines Test	MELD Train	MELD Dev	MELD Test
Emotion	anger	524	85	163	1109	153	345
	disgust	244	26	68	271	22	68
	fear	190	29	36	268	40	50
	joy	1283	123	304	1743	163	402
	neutral	4752	491	1287	4710	470	1256
	sadness	351	62	85	683	111	208
	surprise	1221	151	286	1205	150	281
Sentiment	negative	-	-	-	2945	406	833
	neutral	-	-	-	4710	470	1256
	positive	-	-	-	2334	233	521

Table 4: Emotion and Sentiment distribution in MELD vs. EmotionLines.

in EmotionLines but after viewing the associated video clip, it is correctly re-labeled as *anger* in MELD.

The video of this utterance reveals an angry and frustrated facial expression along with a high vocal pitch, thus helping to recognize its correct emotion. The annotators of EmotionLines had access to the context, but this was not sufficient, as the availability of additional modalities can sometime bring more information for the classification of such instances. These scenarios justify both *context* and *multimodality* to be important aspects for emotion recognition in conversation.

Timestamp alignment. There are many utterances in the subtitles that are grouped within identical timestamps in the subtitle files. In order to find the accurate timestamp for each utterance, we use a transcription alignment tool *Gentle*,[2] which automatically aligns a transcript with the audio by extracting word-level timestamps from the audio (see Table 6). In Table 7, we show the final format of the MELD dataset.

Dyadic MELD. We also provide another version of MELD where all the non-extendable contiguous dyadic sub-dialogues of MELD are extracted. For example, let a three-party dialogue in MELD with speaker ids $1, 2, 3$ have their turns in the following

[2] http://github.com/lowerquality/gentle

order: $[1, 2, 1, 2, 3, 2, 1, 2]$.

From this dialogue sequence, dyadic MELD will have the following sub-dialogues as samples: $[1, 2, 1, 2]$, $[2, 3, 2]$ and $[2, 1, 2]$. However, the reported results in this paper are obtained using only the multiparty variant of MELD.

Utterance	Speaker	MELD	EmotionLines
I'm so sorry!	Chandler	sadness	sadness
Look!	Chandler	surprise	surprise
This guy fell asleep!	Chandler	anger	non-neutral

Table 5: Difference in annotation between EmotionLines and MELD.

3.2 Dataset Exploration

As mentioned before, we use seven emotions for the annotation, i.e., *anger, disgust, fear, joy, neutral, sadness*, and *surprise*, across the training, development, and testing splits (see Table 4). It can be seen that the emotion distribution in the dataset is expectedly non-uniform with the majority emotion being *neutral*. We have also converted these fine-grained emotion labels into more coarse-grained sentiment classes by considering *anger, disgust, fear, sadness* as *negative*, *joy* as *positive*, and *neutral* as *neutral* sentiment-bearing class. *Surprise* is an example of a complex emotion which can be expressed with both positive and negative sentiment. The three annotators who performed the utterance annotation further annotated the *surprise* utterances into either positive or negative sentiment classes. The entire sentiment annotation task reaches a Fleiss' kappa score of 0.91. The distribution of *positive, negative, neutral* sentiment classes is given in Table 4.

Table 8 presents several key statistics of the dataset. The average utterance length – i.e. number of words in an utterance – is nearly the same across training, development, and testing splits. On average, three emotions are present in each dialogue of the dataset. The average duration of an utterance is 3.59 seconds. The emotion shift of a speaker in a dialogue makes emotion recognition task very challenging. We observe that the number of such emotion shifts in successive utterances of a speaker in a dialogue is very frequent: 4003, 427, and 1003 in train/dev/test splits, respectively. Figure 1 shows an example where speaker's emotion changes with time in the dialogue.

Character Distribution. In Figure 3, we present the distributional details of the primary characters in MELD. Figure a and b illustrate the distribution across the emotion and sentiment labels, respectively. Figure c shows the overall coverage of the speakers across the dataset. Multiple infrequent speakers ($< 1\%$ utterances) are grouped as *Others*.

3.3 Related Datasets

Most of the available datasets in multimodal sentiment analysis and emotion recognition are non-conversational. MOSI (Zadeh et al., 2016b), MOSEI (Zadeh et al., 2018), and MOUD (Pérez-Rosas et al., 2013) are such examples that have drawn significant interest from the research community. On the other hand, IEMOCAP and SEMAINE are two popular dyadic conversational datasets where each utterance in a dialogue is labeled by emotion.

The SEMAINE Database is an audiovisual database created for building agents that can engage a person in a sustained and emotional conversation (McKeown et al., 2012). It consists of interactions involving a *human* and an *operator* (either a machine or a person simulating a machine). The dataset contains 150 participants, 959 conversations, each lasting around 5 minutes. A subset of this dataset was used in AVEC 2012's *fully continuous sub-challenge* (Schuller et al., 2012) that requires predictions of four continuous affective dimensions: *arousal, expectancy, power,* and *valence*. The gold annotations are available for every 0.2 second in each video for a total of 95 videos comprising $5,816$ utterances.

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) consists of videos of dyadic conversations among pairs of 10 speakers spanning 10 hours of various dialogue scenarios (Busso et al., 2008). Videos are segmented into utterances with annotations of fine-grained emotion categories: *anger, happiness, sadness, neutral, excitement,* and *frustration*. IEMOCAP also provides continuous attributes: *activation*, *valence*, and *dominance*. These two types of discrete and continuous emotional descriptors facilitate the complementary insights about the emotional expressions of humans and emotional communications between people. The labels in IEMOCAP were annotated by at least three annotators per utterance and self-assessment manikins (SAMs) were also employed to evaluate the corpus (Bradley and Lang, 1994).

3.4 Comparison with MELD

Both resources mentioned above are extensively used in this field of research and contain settings

Utterance	Season	Episode	Incorrect Splits		Corrected Splits	
			Start Time	End Time	Start Time	End Time
Chris says they're closing down the bar.	3	6	00:05:57,023	00:05:59,691	00:05:57,023	00:05:58,734
No way!	3	6	00:05:57,023	00:05:59,691	00:05:58,734	00:05:59,691

Table 6: Example of timestamp alignment using the Gentle alignment tool.

Utterance	Speaker	Emotion	D_ID	U_ID	Season	Episode	StartTime	EndTime
But then who? The waitress I went out with last month?	Joey	surprise	1	0	9	23	00:36:40,364	00:36:42,824
You know? Forget it!	Rachel	sadness	1	1	9	23	00:36:44,368	00:36:46,578

Table 7: MELD dataset format for a dialogue. Notations: D_ID = dialogue ID, U_ID = utterance ID. StartTime and EndTime are in hh:mm:ss,ms format.

that are aligned to the components of MELD. However, MELD is different in terms of both complexity and quantity. Both IEMOCAP and SEMAINE contain dyadic conversations, wherein the dialogues in MELD are multi-party. Multi-party conversations are more challenging compared to dyadic. They provide a flexible setting where multiple speakers can engage. From a research perspective, such availability also demands proposed dialogue models to be scalable towards multiple speakers. MELD also includes more than 13000 emotion labeled utterances, which is nearly double the annotated utterances in IEMOCAP and SEMAINE. Table 9 provides information on the number of available dialogues and their constituent utterances for all three datasets, i.e., IEMOCAP, SEMAINE, and MELD. Table 10 shows the distribution for common emotions as well as highlights a few key statistics of IEMOCAP and MELD.

4 Experiments

4.1 Feature Extraction

We follow Poria et al. (2017) to extract features for each utterance in MELD. For textual features, we initialize each token with pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014) and feed them to a 1D-CNN to extract 100 dimensional textual features. For audio, we use the popular toolkit openSMILE (Eyben et al., 2010), which extracts 6373 dimensional features constituting several low-level descriptors and various statistical functionals of varied vocal and prosodic features. As the audio representation is high dimensional, we employ L2-based feature selection with sparse estimators, such as SVMs, to get a dense representation of the overall audio segment. For the baselines, we do not use visual features, as video-based speaker identification and localization is an open problem. Bimodal features are obtained by concatenating audio and textual features.

MELD Statistics	Train	Dev	Test
# of modalities	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg./Max utterance length	8.0/69	7.9/37	8.2/45
# of dialogues	1039	114	280
# of dialogues dyadic MELD	2560	270	577
# of utterances	9989	1109	2610
# of speakers	260	47	100
Avg. # of utterances per dialogue	9.6	9.7	9.3
Avg. # of emotions per dialogue	3.3	3.3	3.2
Avg./Max # of speakers per dialogue	2.7/9	3.0/8	2.6/8
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Table 8: Dataset Statistics. {a,v,t} = {audio, visual, text}

4.2 Baseline Models

To provide strong benchmarks for MELD, we perform experiments with multiple baselines. Hyperparameter details for each baseline can be found at http://github.com/senticnet/meld.

text-CNN applies CNN to the input utterances without considering the context of the conversation (Kim, 2014). This model represents the simplest baseline which does not leverage context or multimodality in its approach.

bcLSTM is a strong baseline proposed by Poria et al. (2017), which represents context using a bi-directional RNN. It follows a two-step hierarchical process that models uni-modal context first and then bi-modal context features. For unimodal text, a CNN-LSTM model extracts contextual representations for each utterance taking the GloVe em-

Dataset	Type	# dialogues			# utterances		
		train	dev	test	train	dev	test
IEMOCAP	acted	120		31	5810		1623
SEMAINE	acted	58		22	4386		1430
MELD	acted	1039	114	280	9989	1109	2610

Table 9: Comparison among IEMOCAP, SEMAINE, and proposed MELD datasets

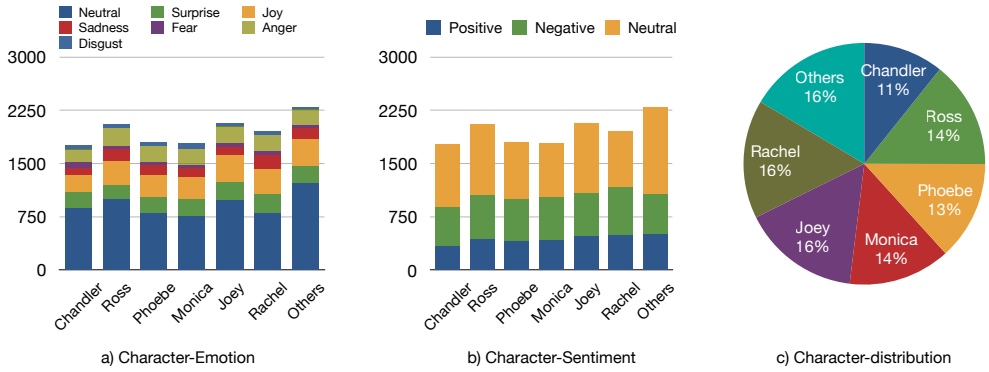

Figure 3: Character distribution across MELD.

Dataset	Emotions						Other Statistics		
	Happy/Joy	Anger	Disgust	Sadness	Surprise	Neutral	Avg. utterence length	#Unique words	Avg. conversation length
IEMOCAP	648	1103	2	1084	107	1708	15.8	3,598	49.2
MELD	2308	1607	361	1002	1636	6436	8.0	10,643	9.6

Table 10: Comparison among IEMOCAP and proposed MELD datasets.

beddings as input. For unimodal audio, an LSTM model gets audio representations for each audio utterance feature vector. Finally, the contextual representations from the unimodal variants are supplied to the bimodal model for classification. bcLSTM does not distinguish among different speakers and models a conversation as a single sequence.

DialogueRNN represents the current state of the art for conversational emotion detection (Majumder et al., 2019). It is a strong baseline with effective mechanisms to model context by tracking individual speaker states throughout the conversation for emotion classification. DialogueRNN is capable of handling multi-party conversation so it can be directly applied on MELD. It employs three stages of gated recurrent units (GRU) (Chung et al., 2014) 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 based on 1) the utterance spoken, 2) the speaker's previous state, and 3) the conversational context 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. Attention mechanism is used on top of the *emotion GRU* to leverage contextual utterances by different speakers at various distances. To analyze the role of multimodal signals, we analyze DialogueRNN and bcLSTM on MELD for both uni and multimodal settings. Training involved usage of class weights to alleviate imbalance issues.

4.3 Results

We provide results for the two tasks of sentiment and emotion classification on MELD. Table 13 shows the performance of sentiment classification by using DialogueRNN, whose multimodal variant achieves the best performance (67.56% F-score) surpassing multimodal bcLSTM (66.68% F-score). Multimodal DialogueRNN also outperforms its unimodal counterparts. However, the improvement due to fusion is about 1.4% higher than the textual modality which suggests the possibility of further improvement through better fusion mechanisms. The textual modality outperforms the audio modality by about 17%, which indicates the importance of spoken language in sentiment analysis. For positive sentiment, audio modality performs poorly. It would be interesting to analyze the clues specific to positive sentiment bearing utterances in MELD that the audio modality could not capture. Future work should aim for enhanced audio feature extraction schemes to improve the classification performance. Table 11 presents the results of the baseline models on MELD emotion classification. The performance on the emotion classes *disgust*, *fear*, and *sadness* are particularly poor. The primary reason for this is the inherent imbalance in the dataset which has fewer training instances for these mentioned emotion classes (see Table 4). We partially tackle this by using class-weights as hyper-parameters.

Yet, the imbalance calls for further improvement for future work to address. We also observe high

Models		Emotions							
		anger	disgust	fear	joy	neutral	sadness	surprise	w-avg.
text-CNN		34.49	8.22	3.74	49.39	74.88	21.05	45.45	55.02
cMKL	text+audio	39.50	16.10	3.75	51.39	72.73	23.95	46.25	55.51
bcLSTM	text	42.06	21.69	7.75	54.31	71.63	26.92	48.15	56.44
	audio	25.85	6.06	2.90	15.74	61.86	14.71	19.34	39.08
	text+audio	43.39	23.66	9.38	54.48	76.67	24.34	51.04	59.25
DialogueRNN	text	40.59	2.04	8.93	50.27	75.75	24.19	49.38	57.03
	audio	35.18	5.13	5.56	13.17	65.57	14.01	20.47	41.79
	text+audio	43.65	7.89	11.68	54.40	77.44	34.59	52.51	**60.25**

Table 11: Test-set weighted F-score results of DialogueRNN for emotion classification in MELD. Note: *w-avg* denotes weighted-average. text-CNN and cMKL: contextual information were not used.

mis-classification rate between the *anger*, *disgust*, and *fear* emotion categories as these emotions have subtle differences among them causing harder disambiguation. Similar to sentiment classification trends, the textual classifier outperforms (57.03% F-score) the audio classifier (41.79% F-score).

Multimodal fusion helps in improving the emotion recognition performance by 3%. However, multimodal classifier performs worse than the textual classifier in classifying sadness. To analyze further, we also run experiments on 5-class emotions by dropping the infrequent *fear* and *disgust* emotions (see Table 12). Not surprisingly, the results improve over the 7-class setting with significantly better performance by the multimodal variant.

Overall, emotion classification performs poorer than sentiment classification. This observation is expected as emotion classification deals with classification with more fine-grained classes.

4.4 Additional Analysis

Role of Context. One of the main purposes of MELD is to train contextual modeling in a conversation for emotion recognition. Table 11 and 13 show that the improvement over the non-contextual model such as text-CNN – which only uses a CNN (see Section 4.1) – is 1.4% to 2.5%.

Inter-speaker influence. One of the important considerations while modeling conversational emotion dynamics is the influence of fellow speakers in the multi-party setting. We analyze this factor by looking at the activation of the attention module on the *global GRU* in DialogueRNN. We observe that in 63% (882/1381) of the correct test predictions, the highest historical attention is given to utterances from different speakers. This significant proportion suggests inter-speaker influence to be an important parameter. Unlike DialogueRNN, bcLSTM does not utilize speaker information while detecting emotion. Table 11 shows that in all the experiments, DialogueRNN outperforms bcLSTM by 1-2% margin. This result supports the claim by Majumder et al. (2019) that speaker-specific modeling of emotion recognition is beneficial as it helps in improving context representation and incorporates important clues such as inter-speaker relations.

	Mode	Emotions					
		ang	joy	neu	sad	surp	w-avg.
bcLSTM	T+A	45.9	52.2	77.9	11.2	49.9	60.6
dRNN*	T	41.7	53.7	77.8	21.2	47.7	60.8
	A	34.1	18.8	66.2	16.0	16.6	44.3
	T+A	48.2	53.2	77.7	20.3	48.5	**61.6**

*dRNN: DialogueRNN, T: text, A: audio

Table 12: Test-set weighted F-score results of DialogueRNN for 5-class emotion classification in MELD. Note: *w-avg* denotes weighted-average. *surp*: surprise emotion.

	Mode	Sentiments			
		pos.	neg.	neu.	w-avg.
	text-CNN	53.23	55.42	74.69	64.25
bcLSTM	T+A	74.68	57.87	60.04	66.68
dRNN*	T	54.35	60.10	74.94	66.10
	A	25.47	45.53	62.33	49.61
	T+A	54.29	58.18	78.40	**67.56**

Table 13: Test set weighted F-score results of DialogueRNN for sentiment classification in MELD.

Emotion shifts. The ability to anticipate the emotion shifts within speakers throughout the course of a dialogue has synergy with better emotion classification. In our results, DialogueRNN achieves a recall of 66% for detecting emotion shifts. However, in the ideal scenario, we would want to detect shift along with the correct emotion class. For this setting, DialogueRNN gets a recall of 36.7%. The deterioration observed is expected as solving both tasks together has a higher complexity. Future methods would need to improve upon their capabilities of detecting shifts to improve the emotion

classification.

Contextual distance. Figure 4 presents the distribution of distances between the target utterance and its second highest attended utterance within the conversation by DialogueRNN in its *emotion GRU*. For the highest attention, the model largely focuses on utterances nearby to the target utterance. However, the dependency on distant utterances increases with the second highest attention. Moreover, it is interesting to see that the dependency exists both towards the historical and the future utterances, thus incentivizing utilization of bi-directional models.

5 Future Directions

Future research using this dataset should focus on improving contextual modeling. Helping models reason about their decisions, exploring emotional influences, and identifying emotion shifts are promising aspects. Another direction is to use visual information available in the raw videos. Identifying face of the speaker in a video where multiple other persons are present is very challenging. This is the case for MELD too as it is a multi-party

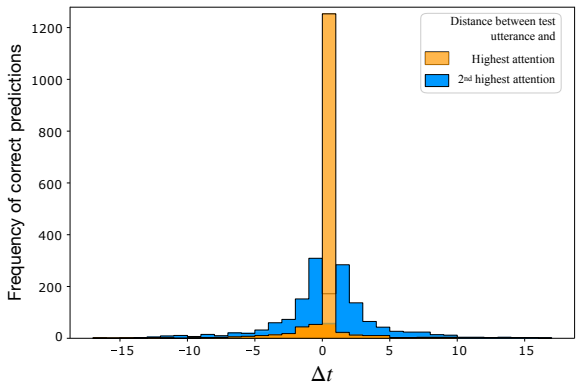

Figure 4: Histogram of Δt = distance between the target and its context utterance based on *emotion GRU* attention scores.

dataset. Enhancements can be made by extracting relevant visual features through processes utilizing audio-visual speaker diarization. Such procedures would enable utilizing a visual modality in the baselines. In our results, audio features do not help significantly. Thus, we believe that it is necessary to improve the feature extraction for these auxiliary modalities in order to improve the performance further.

So far, we have only used concatenation as a feature fusion approach, and showed that it outperforms the unimodal baselines by about 1-3%. We believe there is room for further improvement using other more advanced fusion methods such as MARN (Zadeh et al., 2018).

6 Applications of MELD

MELD has multiple use-cases. It can be used to train emotion classifiers to be further used as emotional receptors in generative dialogue systems. These systems can be used to generate empathetic responses (Zhou et al., 2017). It can also be used for emotion and personality modeling of users in conversations (Li et al., 2016).

By being multimodal, MELD can also be used to train multimodal dialogue systems. Although by itself it is not large enough to train an end-to-end dialogue system (Table 1), the procedures used to create MELD can be adopted to generate a large-scale corpus from any multimodal source such as popular sitcoms. We define *multimodal dialogue system* as a platform where the system has access to the speaker's voice and facial expressions which it exploits to generate responses. Multimodal dialogue systems can be very useful for real time personal assistants such as Siri, Google Assistant where the users can use both voice and text and facial expressions to communicate.

7 Conclusion

In this work, we introduced MELD, a multimodal multi-party conversational emotion recognition dataset. We described the process of building this dataset, and provided results obtained with strong baseline methods applied on this dataset. MELD contains raw videos, audio segments, and transcripts for multimodal processing. Additionally, we also provide the features used in our baseline experiments. We believe this dataset will also be useful as a training corpus for both conversational emotion recognition and multimodal empathetic response generation. Building upon this dataset, future research can explore the design of efficient multimodal fusion algorithms, novel ERC frameworks, as well as the extraction of new features from the audio, visual, and textual modalities.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (grant #1815291), by the John Templeton Foundation (grant #61156), and by DARPA (grant #HR001117S0026-AIDA-FP-045).

References

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR*, abs/1412.3555.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *NAACL*, volume 1, pages 2122–2132.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*, volume 1, pages 994–1003.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Thirty-Third AAAI Conference on Artificial Intelligence*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *ACL (1)*, pages 973–982.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. *arXiv preprint arXiv:1811.00207*.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Amir Zadeh, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016a. Deep constrained local models for facial landmark detection. *arXiv preprint arXiv:1611.08657*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, volume 1, pages 2236–2246.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.