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

Soujanya Poria[‡], Devamanyu Hazarika[Φ], Navonil Majumder[†], Gautam Naik[‡], Rada Mihalcea[ι], Erik Cambria[‡]

[‡]School of Computer Science and Engineering, Nanyang Technological Univerisity, Singapore
[†]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[Φ]School of Computing, National University of Singapore, Singapore
[ι]Computer Science & Engineering, University of Michigan, Ann Arbor, USA

sporia@ntu.edu.sg, hazarika@comp.nus.edu.sg, navo@nlp.cic.ipn.mx, mihalcea@umich.edu, gautam@sentic.net, cambria@ntu.edu.sg

Abstract

Emotion recognition in conversation is a challenging AI task. Recently, it has gained popularity due to its potential applications in many interesting artificial intelligence problems such as empathetic dialogue generation, user behavior understanding etc. To the best of our knowledge, there is no multimodal multi-party conversational dataset available, which contains more than two speakers in a dialogue. On the other hand, although EmotionLines dataset consists of multi-party conversations, it only has dialogues in textual form. To this end, we propose *Multimodal EmotionLines Dataset* (MELD), which we created by enhancing and extending the EmotionLines dataset. MELD contains the same dialogue instances available in EmotionLines, but it also encompasses audio and visual modality along with the text. We have also addressed several problems in EmotionLines and proposed a strong multimodal baseline.

1 Introduction

Multimodal data analysis exploits information from multiple-parallel data channels for decision making. With the rapid growth of AI, multimodal emotion recognition has gained a major research interest, primarily due to its potential applications in many challenging tasks, such as dialogue generation, multimodal interaction etc. A conversational emotion recognition system can be used to generate appropriate responses by analyzing user emotions (Zhou et al., 2017).

Although there are numerous works carried out on multimodal emotion recognition (Poria et al., 2017a; Zadeh et al., 2016a; Wollmer et al., 2013) using audio, visual and text, only a very few actually focus on understanding emotions in conversations. Recently, Hazarika et al. (2018) proposed a multimodal memory network which can recognize emotion in dyadic dialogues. However, their work is limited only to dyadic conversation understanding and thus not scalable to emotion recognition in multi-party conversations having more than two participants. EmotionLines (Chen et al., 2018) is a dataset which contains dialogues from Friends TV series where more than two speakers participate in a dialogue. EmotionLines can be used as a resource for emotion recognition for text only, as it does not include data from other modalities such as visual and audio. At the same time, it should be noted that there is no *multimodal* multi-party conversational dataset available for emotion recognition research. In this work, we have extended, improved, and further developed EmotionLines dataset for the multimodal scenario. In our dataset, not only each dialogue is present in the textual form, but also their corresponding visual and audio counterpart are available.

In a conversation, the participants utter an utterance mostly depending on the context of the conversation. Hence, the emotion expressed by the utterances in a conversation also depend on the context of the conversation. In particular, we can think of conversational context as a set of parameters which influence a person to utter an utterance with an emotion. With the major recent research interests in dialogue systems, studies have been carried out to approach context modeling using different techniques for e.g., using memory networks and RNN (Hazarika et al., 2018; Poria et al., 2017b; Serban et al., 2017). We demonstrate the role of context in Figure 1 where both the speakers change their emotion as the conversation continues dependening on each other's utterance and expressed emotions. Specifically, emotion of utterance 7 in Figure 1 is hard to determine if we dont consider the facial expression. While modeling context in a conversation, such complex inter-speaker relation is one of the major challenges (Hazarika et al., 2018) we encounter. Hazarika et al. (2018) claimed that it is not enough to just use an LSTM or any other network that takes all the previous utterances as input and generates a vector to represent the context. According to them, the model should know the speaker of each utterance and they experimentally showed that this helps in producing better context representation relevant to emotion recognition by means of inter-speaker dependency modeling. Their model dynamically attends to the history of utterances by the same speaker or the other speaker for emotion recognition.

Conversation in its most simplest and natural form is multimodal. We try to rely on others' facial expression, vocal tone, language, gestures etc. while participating in a conversation. It helps us to better understand other participants in the conversation. As far as emotion recognition in a conversation is concerned, multimodality plays a key role. For example, if the spoken language is confusing to perceive the expressed emotion, we often rely on the the vocal tone and facial expression. Hence, in order to create a conversational AI for emotion recognition or other purposes, it is crucial to utilize both the contextual and multimodal information. We believe that MELD will be very useful to meet these two purposes.

As discussed above, emotion recognition in sequential turns has several challenges and classifying short utterances is one of them. Utterances like "yeah", "okay", "no" can express different emotions depending on the context and discourse of the dialogue. The emotion change and emotion flow in the sequence of turns in a dialogue make accurate context modeling a difficult task. In this dataset, as we have access to the multimodal data sources for each dialogue, we hypothesize that it will improve the context representation, supplement missing or misleading information from other moralities thus benefiting the overall emotion recognition performance. Access to multiple modalities is also useful in classifying utterances such as "yeah", "okay", "I'll". These utterances do not express any explicit emotion by themselves, but the speaker's facial expressions or intonation in speech could carry important clues for classifying such utterances as *non-neutral*. This dataset can also be used in a multimodal affective dialogue system.

IEMOCAP (Busso et al., 2008), SEMAINE (Schuller et al., 2012) are multimodal conversational datasets which contain emotion label for each utterance. However, these datasets are dyadic in nature, which justifies the importance of our *Multimodal-EmotionLines* dataset. The other publicly available multimodal emotion and sentiment recognition datasets are MOSEI (Zadeh et al., 2018b), MOSI (Zadeh et al., 2016b), MOUD (Pérez-Rosas et al., 2013). However, none of those datasets is conversational.

2 Related Datasets

Most of the available datasets in multimodal sentiment analysis and emotion recognition are not conversational. MOSI (Hazarika et al., 2018; Zadeh et al., 2017), MOSEI (Zadeh et al., 2018a,b), MOUD (Poria et al., 2016) are such non-conversational datasets which have drawn major research interests. On the other hand, IEMOCAP and SEMAINE are the dyadic conversational datasets where each utterance in a dialogue is labeled by emotion. As these two datasets are similar to MELD, we limit the scope of this section to only IEMOCAP and SEMAINE.

The SEMAINE Database was developed by McKeown et al. (2012). It is a large audiovisual database created for building agents that can engage a person in a sustained and emotional conversation using Sensitive Artificial Listener (SAL) (Douglas-Cowie et al., 2008) paradigm. SAL is an interaction involving two parties: a 'human' and an 'operator' (either a machine or a person simulating a machine). The interaction is based on two qualities: one is low sensitivity to preceding verbal context (the words the user used that do not dictate whether to continue the conversation) and the second is conduciveness (response to a phrase by continuing the conversation). There were 150 participants, 959 conversations, each lasting 5 minutes. There were 6-8 annotators per clip, who eventually traced 5 affective dimensions and 27 associated categories. For the recordings, the participants were asked to talk in turn to four emotionally stereotyped characters. The characters are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is sad and depressive. Videos were recorded

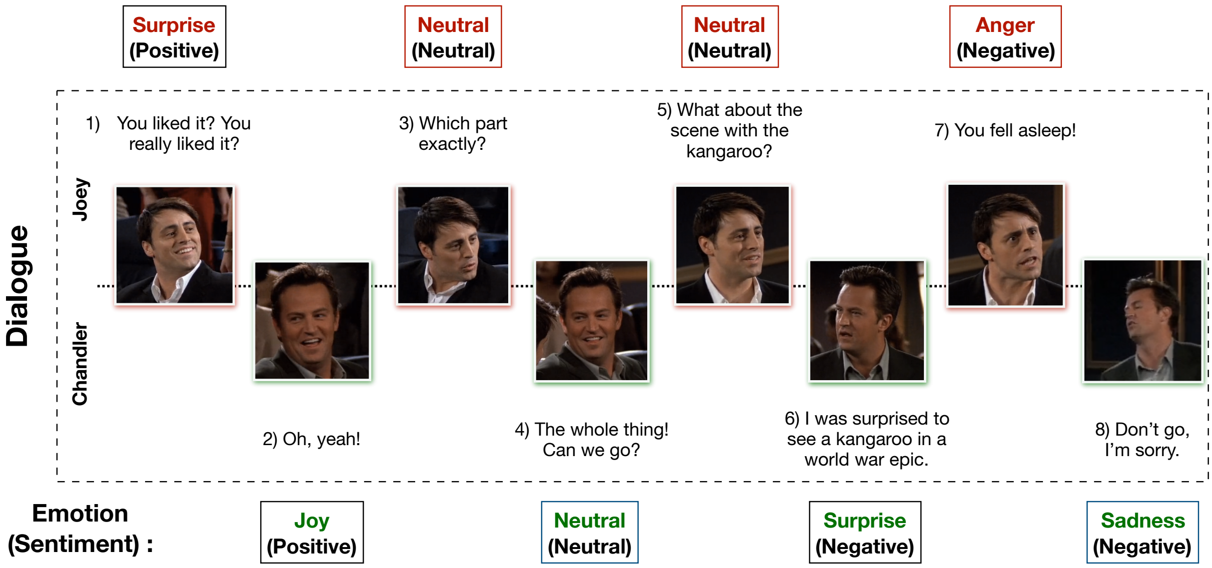

Figure 1: Emotion shift of speakers in a dialogue in comparison with speaker's previous emotion. Red and blue colors are used to show the emotion shift of Chandler and Monica respectively.

at 49.979 frames per second at a spatial resolution of 780 x 580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample. To accommodate research in audio-visual fusion, the audio and video signals were synchronized with an accuracy of 25 micro-seconds.

Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) IEMOCAP dataset was developed by Busso et al. (2008). 10 actors were asked to record their facial expressions in front of cameras. Facial markers, head and hand gesture trackers were placed in order to collect the facial expressions, head and hand gestures. In particular, the dataset contains a total 10 hours of recording of dyadic sessions. Each recording of the dataset expresses either of these emotions - happiness, anger, sadness, frustration and neutral state. The recorded dyadic sessions were later manually segmented at the utterance level which is defined as continuous segments when one of the actors was actively speaking. The acting was based on some scripts, hence it was easy to segment the dialogues for utterance detection in the textual part of the recordings. Busso et al. (2008) used two famous emotion taxonomies in order to manually label the dataset in utterance level: discrete categorical based annotations (i.e., labels such as happiness, anger, and sadness), and continuous attribute based annotations (i.e., activation, valence, and dominance). To assess the emotion categories of the recordings, six human annotators were appointed. Having two different annotation schemes can provide complementary information in human-machine interaction system. The evaluation sessions were organized so that three different annotators assessed each utterance. Self-assessment manikins (SAMs) were also employed to evaluate the corpus in terms of the attributes valence [1-negative, 5-positive], activation [1-calm, 5-excited], and dominance [1-weak, 5-strong]. Two more human annotators were asked to estimate the emotional content in recordings using the SAM system. These two types of emotional descriptors facilitate the complementary insights about the emotional expressions of humans, emotional communications between people which can further potentially help to develop a better human-machine interfaces by automatically recognizing and synthesizing the emotional cues expressed by humans.

MELD is superior to these two datasets in terms of both complexity and quantity. Both IEMOCAP and SEMAINE contain only dyadic conversations wherein the dialogues in MELD are multiparty. Multiparty conversations are more challenging in comparison to dyadic. MELD has more than 13000 emotion labeled utterances which is almost double of the annotated utterances present in both IEMOCAP and SEMAINE. In Table 10 we present a comparison among MELD, IEMOCAP and SEMAINE. We discuss this comparison in more detail in *Comparison with the Related Datasets* section.

3 EmotionLines Dataset

EmotionLines dataset was developed by Chen et al. (2018). This dataset contains dialogues from sitcom Friends, where each dialogue contains utterances from multiple speakers who participated. Chen et al. (2018) crawled the dialogues from each episode and grouped the them into four groups ([5, 9], [10, 14], [15, 19], and [20, 24]) based on the number of utterances present in the dialogues. Finally, 250 dialogues were sampled randomly from each of these groups which finally contributed to a dataset consisting of 1000 dialogues.

3.1 Annotation

The utterances in each dialogue were annotated with the most appropriate emotion category. Chen et al. (2018) considered Ekman's six emotions, i.e., *Joy, Sadness, Fear, Anger, Surprise, and Disgust* as annotation labels. This annotation list was extended with an extra emotion label *Neutral*. Chen et al. (2018) used Amazon Mechanical Turk (AMT) to annotate the utterances. The authors used five Mturkers for the annotation. Majority voting scheme was applied in order to select a final emotion label for each utterance. The overall kappa score of this annotation process was 0.34.

4 Multimodal EmotionLines Dataset (MELD)

We further extended the EmotionLines dataset into a multimodal dataset. Below are the steps that were taken to construct the dataset:

1. The first step deals with finding the timestamp of every utterance in each of the dialogues present in the EmotionLines dataset. To accomplish this, we crawled through the subtitle files of all the episodes which contains the beginning and the end timestamp of the utterances. This process enabled us to obtain season ID, episode ID, and timestamp of each utterance in the episode. We put two constraints whilst obtaining the timestamps:

   (a) timestamps of the utterances in a dialogue must be in increasing order,

   (b) all the utterances in a dialogue have to belong to the same episode and scene.

   Constraining with these two conditions revealed that in EmotionLines, a few dialogues consist of multiple natural dialogues. We filtered out those cases from the dataset. One such example from EmotionLines is shown in Table 2. The dialogue in Table 2 contains two natural dialogues from episode 4 and 20 of season 6 and 5 respectively. Because of this error correction step, in our case, we have the different number of dialogues as compare to the EmotionLines.

2. We then asked three annotators to label each utterance in a dialogue. Majority voting was applied to decide the final label of the utterances. A few utterances were removed which did not have majority annotators' agreement.

3. After obtaining the timestamp of each utterance, we extracted their corresponding audio-visual clips from the source episode. Separately, we also took out the audio content from those video clips. Finally, the dataset contains visual, audio, and textual modality for each dialogue.

	#Dialogues	
	EmotionLines	MELD
Train	720	1039
Dev	80	114
Test	200	280

Table 1: Comparison between original EmotionLines and multimodal EmotionLines dataset (MELD).

4.1 Dataset Re-annotation

The utterances in the original EmotionLines dataset were annotated by only looking at the textual part of the utterances. However, as we have focused to develop a multimodal version of the EmotionLines dataset, we re-annotated all the utterances by asking three annotators to also look at the available video clip of the utterances. Majority voting technique was taken to obtain final label out of the three annotations for each utterance. The Fleiss's kappa score of this annotation process was 0.43 which is higher than the kappa of the original EmotionLines annotation. We removed those utterances for which we could not find majority agreement among the annotators. Table 4 shows the label wise comparison between the original EmotionLines and re-annotated Multimodal

Episode	Utterance	Speaker	Emotion	Sentiment
S6.E4	Hey Estelle, listen	Joey	neutral	neutral
	Well! Well! Well! Joey Tribbiani! So you came back huh? They	Estelle	surprise	positive
	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: An incorrect dialogue in EmotionLines where utterances from different episodes are present. First 6 utterances in this dialogue have been taken from episode 4 of season 6. The last 3 utterances in red color are from episode 20 of season 5.

EmotionLines dataset i.e., MELD. For most of the utterances, annotations in MELD match with the original EmotionLines' annotations. When asked, annotators confirmed that the video clips of the utterances helped them in the annotation. One such utterance is *"This guy fell asleep!"* (as shown in Table 3). This utterance has been labeled as *non-neutral* in EmotionLines but thanks to the available video clip, it has been labeled as *anger* in MELD. Manually looking at the video clip of this utterance reveals that a very angry and frustrated facial expression along with a high vocal tone are key to recognize its' correct emotion. We think that the surrounding contextual utterances of it were not sufficient enough for the EmotionLines' annotators to label this utterance correctly. To this end, this example justifies that both context and multimodality are important aspects for emotion recognition in dialogue or conversation in general.

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

Table 3: EmotionLines vs MELD: difference in the annotation. Non-neutral signifies the case where annotators agreed that the emotion expressed by the utterance is not neutral but they could not reach an agreement about the correct emotion label.

4.2 Dataset Pruning

There were many utterances in the subtitles which are grouped within identical timestamps in the subtitle flies. In order to find the accurate timestamp for each utterance, we used a transcription alignment tool *Gentle* [1], which automatically aligns a transcription with the audio by extracting word-level timestamps from the audio (Table 5). In Table 6, we show the format of MELD dataset.

	EmotionLines			MELD		
Emotion	Train	Dev	Test	Train	Dev	Test
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

Table 4: Emotion distribution in the dataset.

4.3 Dataset Exploration

The number of emotions labels in this dataset is seven, i.e., anger, disgust, fear, joy, neutral, sadness, and surprise. We present the emotion distribution in training, development, and test datasets in Table 7. It can be seen that the emotion distribution in the dataset is not uniform and the majority of the utterances are labeled as *neutral*. The dataset is publicly available for research at https://affective-meld.github.io/.

We have also converted these fine grained emotion labels into more coarse grained sentiment classes by considering *anger, disgust, fear, sadness* as *negative* and *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. Three annotators who were employed for utterance annotation were further asked to annotate

[1] https://github.com/lowerquality/gentle

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 5: Example of dataset pruning using Gentle alignment tool.

Utterance	Speaker	Emotion	U_ID	Season	E_ID	StartTime	EndTime
But then who? The waitress I went out with last month?	Joey	surprise	0	9	23	00:36:40,364	00:36:42,824
You know? Forget it!	Rachel	sadness	1	9	23	00:36:44,368	00:36:46,578
No-no-no-no, no! Who, who were you talking about?	Joey	surprise	2	9	23	00:36:44,368	00:36:49,122
No, I-I-I-I don't, I actually don't know	Rachel	fear	3	9	23	00:36:49,290	00:36:51,791
Ok!	Joey	neutral	4	9	23	00:36:52,376	00:36:53,543
All right, well...	Joey	neutral	5	9	23	00:36:53,545	00:36:55,000
I'm gonna see if I can get a room for the night and I'll...	Joey	neutral	6	9	23	00:36:54,587	00:36:58,000
I'll see you later!	Joey	neutral	7	9	23	00:36:57,506	00:36:59,425
Yeah, sure!	Rachel	neutral	8	9	23	00:36:59,425	00:37:01,439

Table 6: MELD dataset format. Notations: U_ID = utterance ID, E_ID = episode ID. StartTime and EndTime are in hh:mm:ss,ms format.

the surprise emotion bearing utterances into either positive and negative sentiment classes. This whole sentiment annotation task had an Fleiss' kappa score of 0.91. The distribution of *positive, negative, neutral* sentiment classes is given in Table 8.

Table 9 presents several key statistics of the dataset. We can see that the average utterance length, in terms of the number of words present in an utterance, is almost the same across training, development, and test datasets. On average, three emotions are present in a dialogue of the dataset. The average duration of an utterance is 3.59 seconds. Emotion shift of a speaker in a dialogue makes emotion recognition task very challenging. We observe that the number of such emotion shift in successive utterances of a speaker in a dialogue is very frequent in the dataset – 4003, 427, and 1003 in training, development, and test datasets respectively. Figure 1 shows an example where speaker's emotion changes with time in the dialogue.

4.4 Comparison with the Related Datasets

In this section, we compare our proposed MELD dataset with other databases. Particularly, we select two datasets, IEMOCAP[2] (Busso et al., 2008) and SEMAINE[3] (Schuller et al., 2012), that are extensively used in this field of research and contain settings which are aligned to the components of MELD.

Both IEMOCAP and SEMAINE are dyadic conversational databases. IEMOCAP contain annotations in both categorical and continuous dimensions comprising of emotional categories: *Anger, Happiness, Sadness, Neutral, Excitement, Fear, Surprise, Disgust, Frustration, and Others* and continuous emotional dimensions: *valence, arousal, and dominance*. Both the annotations are done at utterance level involving multiple annotators. In contrast, SEMAINE database contains annotations only in continuous affective dimensions that include, *Valence, Activation/Arousal, Power/Dominance, Anticipation/Expectation, Intensity, Fear, Anger, Happiness, Sadness, Disgust, Contempt,* and *Amusement*. Here, annotations are provided at a finer granularity, where, labels exist at a gap of 0.2 seconds for each conversational video.

Table 10 provides information on the number of available dialogues and their constituent utterances for all three datasets, i.e., IEMOCAP, SEMAINE, and MELD. As seen in the table, MELD contains the largest size of dialogues (and utterances) which is significantly more than the other two. Figure 2 also indicates this trend for common emotions between IEMOCAP and MELD. Except for *sadness*, MELD contains a higher amount of

[2] https://sail.usc.edu/iemocap/
[3] https://sspnet.eu/avec2012/

instances pertaining to the respective emotional categories. The extremeness of available *neutral* utterances in MELD emulates real-life conversation trends where the prevailing emotion is generally *neutral*. Another key difference for MELD is that it contains multi-party dialogues whereas IEMOCAP and SEMAINE are datasets comprising dyadic interactions only. This provides a natural setting for dialogues where multiple speakers can engage and demands proposed dialogue models to be scalable towards multiple speakers.

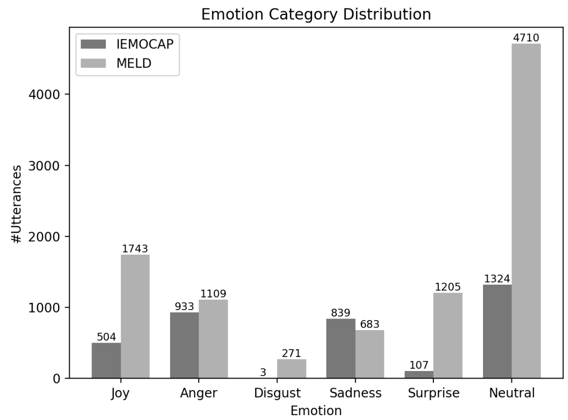

Figure 2: Comparison between the distribution of common emotions between training splits of IEMOCAP and MELD dataset.

5 Strong Baseline

5.1 Unimodal Feature Extraction

In this section, we discuss the method of feature extraction for three different modalities: audio, video, and text. We have followed the contextual multimodal sentiment analysis approach, proposed by Poria et al. (2017b) to get the baseline results on MELD.

	No. of Utterances		
Emotion	Train	Dev	Test
anger	1109	153	345
disgust	271	22	68
fear	268	40	50
joy	1743	163	402
neutral	4710	470	1256
sadness	683	111	208
surprise	1205	150	281

Table 7: Emotion distribution in the dataset.

	No. of Utterances		
Sentiment Category	Train	Dev	Test
negative	2945	406	833
neutral	4710	470	1256
positive	2334	233	521

Table 8: Coarse sentiment distribution in the dataset.

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. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Table 9: Dataset Statistics. {a,v,t} = {audio, visual, text}

5.1.1 Textual Feature Extraction

The textual data is obtained from the transcripts of the videos. We apply a deep Convolutional Neural Networks (CNN) (Karpathy et al., 2014) on each utterance to extract textual features. Each utterance in the text is represented as an array of pre-trained 300-dimensional `Glove` vectors (Pennington et al., 2014). Further, the utterances are truncated or padded with null vectors to have exactly 50 words.

Next, these utterances as an array of vectors are passed through two different convolutional layers; the first layer having two filters of size 3 and 4 respectively with 50 feature maps, each and the second layer has a filter of size 2 with 100 feature maps. Each convolutional layer is followed by a max-pooling layer with window 2×2.

The output of the second max-pooling layer is fed to a fully-connected layer with 500 neurons with a rectified linear unit (ReLU) (Teh and Hinton, 2001) activation, followed by softmax output. The output of the penultimate fully-connected layer is used as the textual feature. The translation of convolution filter over makes the CNN learn abstract features and with each subsequent layer, the context of the features expands further.

5.1.2 Audio Feature Extraction

The audio feature extraction process is performed at 30 Hz frame rate with 100 ms sliding window.

Dataset	# dialogues			# utterances		
	train	dev	test	train	dev	test
IEMOCAP	120		31	5810		1623
SEMAINE	63		32	4368		1430
MELD	1039	114	280	9989	1109	2610

Table 10: Comparison among MELD, IEMOCAP and SEMAINE datasets

We use openSMILE (Eyben et al., 2010), which is capable of automatic pitch and voice intensity extraction, for audio feature extraction. Prior to feature extraction, audio signals are processed with voice intensity thresholding and voice normalization. Specifically, we use Z-standardization for voice normalization. In order to filter out audio segments without the voice, we threshold voice intensity. OpenSMILE is used to perform both these steps. Using openSMILE we extract several Low-Level Descriptors (LLD) (e.g., pitch, voice intensity) and various statistical functionals of them (e.g., amplitude mean, arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, and linear regression slope). "IS13-ComParE" configuration file of openSMILE is used to for our purposes. Finally, we extracted total 6373 features from each input audio segment.

5.2 Context Modeling

Utterances in the videos are semantically dependent on each other. In other words, the complete meaning of an utterance may be determined by taking preceding utterances into consideration. We call this the context of an utterance. Following (Poria et al., 2017b), we use RNN, specifically, GRU[4] to model semantic dependency among the utterances in a video. Let the following items represent unimodal features:

$$f_A \in \mathbb{R}^{N \times d_A} \quad \text{(acoustic features)},$$
$$f_T \in \mathbb{R}^{N \times d_T} \quad \text{(textual features)},$$

where N = maximum number of utterances in a video. We pad the shorter videos with dummy utterances represented by null vectors of corresponding length. For each modality, we feed the unimodal utterance features f_m (where $m \in \{A, T\}$) (discussed in 5.1) of a video to GRU_m with output size D_m, which is defined as

$$\begin{aligned}
z_m &= \sigma(f_{mt} U^{mz} + s_{m(t-1)} W^{mz}),\\
r_m &= \sigma(f_{mt} U^{mr} + s_{m(t-1)} W^{mr}),\\
h_{mt} &= \tanh(f_{mt} U^{mh} + (s_{m(t-1)} * r_m) W^{mh}),\\
F_{mt} &= \tanh(h_{mt} U^{mx} + u^{mx}),\\
s_{mt} &= (1 - z_m) * F_{mt} + z_m * s_{m(t-1)},
\end{aligned}$$

where $U^{mz} \in \mathbb{R}^{d_m \times D_m}$, $W^{mz} \in \mathbb{R}^{D_m \times D_m}$, $U^{mr} \in \mathbb{R}^{d_m \times D_m}$, $W^{mr} \in \mathbb{R}^{D_m \times D_m}$, $U^{mh} \in \mathbb{R}^{d_m \times D_m}$, $W^{mh} \in \mathbb{R}^{D_m \times D_m}$, $U^{mx} \in \mathbb{R}^{d_m \times D_m}$, $u^{mx} \in \mathbb{R}^{D_m}$, $z_m \in \mathbb{R}^{D_m}$, $r_m \in \mathbb{R}^{D_m}$, $h_{mt} \in \mathbb{R}^{D_m}$, $F_{mt} \in \mathbb{R}^{D_m}$, and $s_{mt} \in \mathbb{R}^{D_m}$. This yields hidden outputs F_{mt} as context-aware unimodal features for each modality. Hence, we define $F_m = GRU_m(f_m)$, where $F_m \in \mathbb{R}^{N \times D_m}$. Thus, the context-aware unimodal features can be defined as

$$\begin{aligned}
F_A &= GRU_A(f_A),\\
F_T &= GRU_T(f_T).
\end{aligned}$$

5.3 Fusion

We then fuse F_A, F_T to a multimodal feature space. In order to get the fused representation of the modalities, F_{AT}, we simply concatenated F_A and F_T by following Poria et al. (2017b). The main reason for choosing concatenation based fusion is because it is very simple to implement yet an effective fusion approach. Use of complex state-of-the-art fusion methods such as *Tensor Fusion*Zadeh et al. (2017) are left for future work.

$$F_{AT} = [F_A; F_T]$$

Finally f_{AT} was fed to contextual GRU i.e., GRU_{AT} which incorporates the contextual information contributed by the utterances.

5.4 Classification and Training

The training of this network is performed using categorical cross-entropy on each utterance's softmax output per dialogue, i.e.,

$$loss = -\frac{1}{(\sum_{i=1}^{M} L_i)} \sum_{i=1}^{M} \sum_{j=1}^{L_i} \sum_{c=1}^{C} y_{i,c}^{j} \log_2(\hat{y}_{i,c}^{j}),$$

where M = total number of dialogues in the dataset, L_i = number of utterances for i^{th} dialogue, $y_{i,c}^{j}$ = original output of class c, and $\hat{y}_{i,c}^{j}$ = predicted output for j^{th} utterance of i^{th} dialogue.

[4] LSTM does not perform well

As a regularization method, dropout between the GRU cell and dense layer is introduced to avoid overfitting. We used Adam (Kingma and Ba, 2014) as an optimizer.

5.5 Baseline Results

In Table 11 and 12, we show the baseline results following the method explained in Section *Strong Baseline*. As it can be seen, multimodality outperforms the text and audio modality. However, the improvement due to the fusion is not very higher than the textual modality which suggests the need for a better fusion mechanism. We left that for the future work. Overall, the audio modality performs the worst. For positive sentiment category, audio modality could only produce 2.56% accuracy. It would be interesting to analyze the clues specific to positive emotion bearing utterances in MELD that audio modality could not capture. In future, we will use more advance state-of-the-art audio feature extractor in order to improve the classification performance.

Modality	Sentiments			
	positive	negative	neutral	w-avg.
text-CNN	53.23	55.42	74.69	64.25
text	65.12	82.66	90.50	82.91
audio	2.56	71.39	83.45	63.08
text+audio	67.27	81.64	91.67	83.60

Table 11: Test-set F-score results of contextual biLSTM for sentiment classification in MELD. Note: *w-avg* denotes weighted-average.

In the case of emotion classification, the performance is poor to classify disgust, fear and sadness emotions. We think, this has happened as the number of training instances for disgust, fear and sadness are very low as shown in Table 7. Nevertheless, this performance acts as a baseline and future works should aim at outperforming this baseline. We observed high mis-classification rate for anger, disgust and fear emotion categories since these emotions have a very subtle difference among them which causing harder disambiguation. Overall, emotion classification results are worse than that of sentiment classification. This observation is expected as emotion classification deals with classification into more classes. Multimodal fusion helps in improving emotion recognition performance by a mere 0.41%. However, multimodal classifier performs worse than textual classifier in classifying surprise and sadness emotions. The number of utterances labeled with disgust and fear are only 68 and 50 respectively. Hence, the difference in results of the textual and multimodal classifiers for these two emotions are not statistically significant.

Role of Context One of the main purpose of MELD is to build an AI that utilizes context in a conversation for emotion recognition. We have discussed in the introduction on the importance of contextual information in conversation. Table 11 and 12 show that the improvement over non-contextual models for e.g., text-CNN which only uses a CNN (see Section 5.1.1) is 16% to 19%. We believe this is a very interesting result which should open the door of more research on modeling context.

6 Future Directions

There are a number of interesting future directions of this work. First, the proposed baseline does not consider the presence of multiple speakers in a conversation. We think that the consideration of speaker modeling can improve the performance of emotion recognition. The other future directions include extraction of visual features and use of this dataset for empathetic dialogue generation. As a part of the dataset, we have released the raw videos and audios which will facilitate the feature extraction process.

6.1 Applications of this dataset

The other use cases of this dataset are as follows:

- As we discussed above, this dataset is useful to train a conversational emotion recognition classifier which can be plugged into any dialogue system to generate empathetic responses similar to Zhou et al. (2017). For example, this dataset can be used for emotion modeling of the users in Twitter persona dataset Li et al. (2016). As this dataset is multimodal, it is also possible to integrate it with a multimodal dialogue system.

- This dataset should not be used to train an end-to-end dialogue system because of its size (see Table 1). The training set of this dataset contains only 10,643 utterances, which is not enough to train a well performing dialogue system. However, the mechanism of constructing this dataset can be easily applied to develop a *multimodal* dialogue

Modality	Emotions							
	anger	disgust	fear	joy	neutral	sadness	surprise	w-avg.
text-CNN	31.52	16.05	9.02	50.57	75.32	16.47	47.58	55.12
text	63.12	6.02	7.98	56.26	90.01	39.75	63.23	70.8
audio	42.37	0.0	0.0	46.34	83.26	23.78	32.45	57.6
text+audio	67.23	2.50	3.12	58.28	90.60	37.34	62.39	71.27

Table 12: Test-set F-score results of contextual biLSTM for emotion classification in MELD. Note: *w-avg* denotes weighted-average.

dataset based on Friends or any other TV series such as Breaking Bad. We define *multimodal dialogue system* as a platform where the system has access to the speaker's voice, the facial expression 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 to communicate with the assistant.

7 Conclusion

In this work, we propose a multimodal multiparty conversational emotion recognition dataset called MELD. MELD has been developed based on the original EmotionLines dataset. We also provide solid baseline results on MELD. This dataset is publicly available [5] and contains the raw videos and audios which will be useful to extract new audio and visual features. Along with these, we have also released the features used in our baseline experiments. We think this dataset will also be useful as a training corpus for multimodal emphatic response generation. MELD has a strong potential to help conversation understanding research. Future works on this dataset should focus on extracting new features and outperforming the baseline results as presented in this paper.

Acknowledgments

We are thankful to the authors of the original EmotionLines paper Chen et al. (2018).

References

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*.

Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and DKJ Heylen. 2008. The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. *LREC Workshop on Corpora for Research on Emotion and Affect*.

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2122–2132.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

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

[5] https://affective-meld.github.io/

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 *Proceedings of the 2014 conference on empirical methods in natural language processing (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, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

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

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *ICDM*, pages 439–448, Barcelona.

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.

Vee Teh and Geoffrey E Hinton. 2001. Rate-coded restricted boltzmann machines for face recognition. In *Advances in neural information processing system*, volume 13, pages 908–914.

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, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1114–1125.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018a. Multi-attention recurrent network for human communication comprehension. In *AAAI*, pages 5642–5649.

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.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246.

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*.