Multimodal EmotionLines: 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, gelbukh@gelbukh.com, 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 EmotionLines dataset. MELD contains the same dialogue instances available in EmotionLines, but it also encompasses audio and visual modality along with 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, only a very few actually focus on understanding emotions in conversations. Recently, Hazarika et al. (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 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.

Emotion recognition in sequential turns has several challenges and context understanding 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 modeling 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 to develop a multimodal affective dialogue system.

IEMOCAP (Busso et al., 2008a), AVEC 2012 (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., 2018), MOSI (Zadeh et al., 2016), MOUD (Pérez-Rosas, Mihalcea, and Morency, 2013). However, none of those datasets is conversational.

2 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 1000 dialogues.

2.1 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.

3 Multimodal EmotionLines Dataset

We have further extended the EmotionLines dataset into a multimodal dataset. This process is explained below:

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 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 (shown in Table 3 and 4) from the dataset. One such example from EmotionLines is shown in Table 3. The dialogue in Table 3 contains two natural dialogues from episode 8 and 9 which we correct in Table 4. Hence, in our case, we have the different number of dialogues as compare to the EmotionLines.

2. 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).

3.1 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 of each of those utterances, we used a transcription alignment tool *Gentle*. In particular, *Gentle* [1] automatically aligns a transcription with the audio and finds out the corresponding timestamp of the transcription in the audio (Table 2).

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

			Incorrect Splits		Corrected Splits	
Utterance	Season	Episode	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 2: Example of dataset pruning using Gentle alignment tool

Episode	Utterance	Speaker	Emotion	Sentiment
6.4	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! You've always been my agent!	Joey	surprise	negative
	Really?!	Estelle	surprise	positive
	Yeah!	Joey	joy	positive
	Oh well, no harm, no foul.	Estelle	neutral	neutral
5.20	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 3: 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.

3.2 Dataset Exploration

The number of emotions labels in this dataset is seven, i.e., anger, disgust, fear, joy, neutral, sadness, and surprise. We show the emotion distribution in training, development, and test datasets in Table 5. It can be seen that the emotion distribution in the dataset is not uniform and majority of the utterances are labeled as *neutral*.

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. Since, *surprise* is an example of a complex emotion which can be expressed with both positive and negative sentiment, we did not further generalize this emotion class. The distribution of *positive, negative, neutral, and surprise* emotion classes is given in Table 6.

Table 7 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.

3.3 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., 2008b) 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 8 provides information on the number of available dialogues and their constituent utter-

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

Utterance	Speaker	Emotion	U_ID	Season	E_ID	StartTime	EndTime
		Dialogue 1					
also I was the point person on my company's transition from the KL-5 to GR-6 system.	Chandler	neutral	0	8	21	00:16:16,059	00:16:21,731
You must've had your hands full.	Interviewer	neutral	1	8	21	00:16:21,940	00:16:23,442
That I did. That I did.	Chandler	neutral	2	8	21	00:16:23,442	00:16:26,389
So let's talk a little bit about your duties.	Interviewer	neutral	3	8	21	00:16:26,820	00:16:29,572
My duties? All right.	Chandler	surprise	4	8	21	00:16:34,452	00:16:40,917
Now you'll be heading a whole division, so you'll have a lot of duties.	Interviewer	neutral	5	8	21	00:16:41,126	00:16:44,337
I see.	Chandler	neutral	6	8	21	00:16:48,800	00:16:51,886
But there'll be perhaps 30 people under yo u so you can dump a certain amount on them.	Interviewer	neutral	7	8	21	00:16:48,800	00:16:54,514
Good to know.	Chandler	neutral	8	8	21	00:16:59,477	00:17:00,478
We can go into detail	Interviewer	neutral	9	8	21	00:17:00,478	00:17:02,719
No don't I beg of you!	Chandler	fear	10	8	21	00:17:02,856	00:17:04,858
All right then, we'll have a definite answer for you on Monday, but I think I can say with some confidence, you'll fit in well here.	Interviewer	neutral	11	8	21	00:17:05,025	00:17:13,324
Really?!	Chandler	surprise	12	8	21	00:17:13,491	00:17:16,536
Absolutely. You can relax	Interviewer	neutral	13	8	21	00:17:17,579	00:17:20,707
		Dialogue 2					
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 4: Splitting an incorrect dialogue (shown in Table 3) in EmotionLines into two correct dialogues. Notations: U_ID = utterance ID, E_ID = episode ID

	No. of Utterances		
Emotion	Train	Dev	Test
anger	1110	154	345
disgust	271	22	68
fear	268	40	51
joy	1748	163	405
neutral	4726	471	1261
sadness	686	111	208
surprise	1207	150	282

Table 5: Emotion distribution in the dataset

	No. of Utterances		
Sentiment Category	Train	Dev	Test
negative	2335	327	672
neutral	4726	471	1261
positive	1748	163	405
surprise	1207	150	282

Table 6: Coarse sentiment distribution in the dataset

ances 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 higher amount of 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.

4 Strong Baseline

4.1 Unimodal Feature Extraction

In this section, we discuss the method of feature extraction for three different modalities: audio, video, and text.

4.1.1 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 `word2vec` vec-

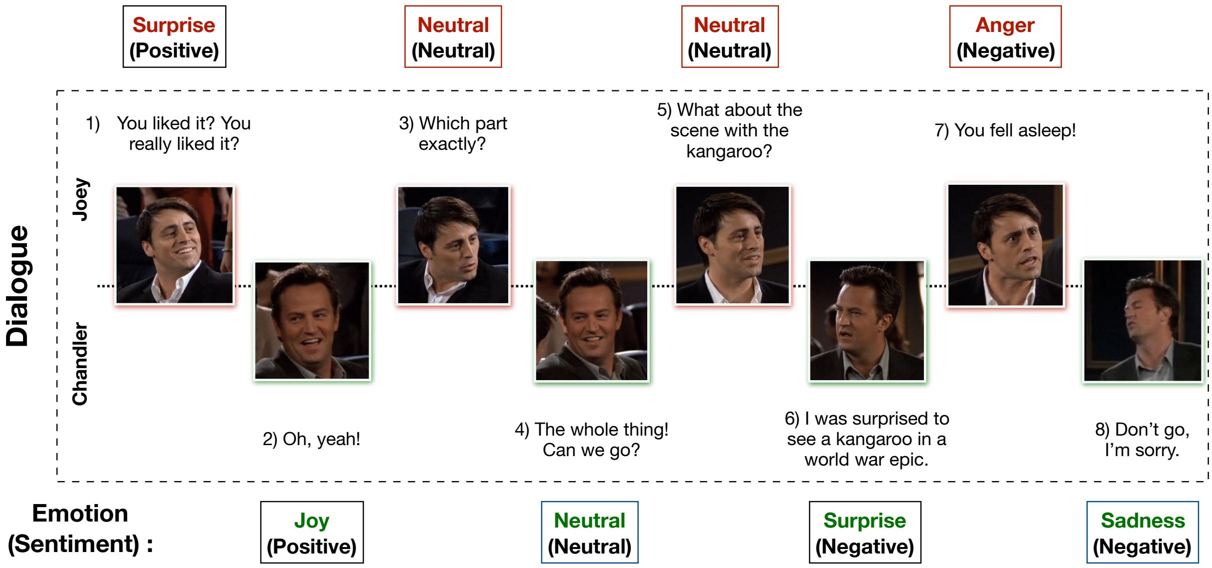

Figure 1: Emotional and sentimental shift of speakers in a dialogue in comparison with speaker's previous states.

Statistics	Train	Dev	Test
# 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	10016	1111	2620
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Table 7: Dataset Statistics

Dataset	# dialogues			# utterances		
	train	dev	test	train	dev	test
IEMOCAP	120		31	5810		1623
SEMAINE	63		32	4368		1430
MELD	1039	114	280	10463	2384	4361

Table 8: Comparison among MELD, IEMOCAP and SEMAINE

tors (Mikolov et al., 2013). Further, the utterances are truncated or padded with null vectors to have exactly 50 words.

Next, these utterances as array of vectors are passed through two different convolutional layers; 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.

4.1.2 Audio Feature Extraction

The audio feature extraction process is performed at 30 Hz frame rate with 100 ms sliding window. We use openSMILE (Eyben, Wöllmer, and Schuller, 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 6392 features from each input audio segment.

4.2 Context Modeling

Utterances in the videos are semantically dependent on each other. In other words, 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., 2017), 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 4.1) 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 multimodal features can be defined as

$$F_A = GRU_A(f_A),$$
$$F_T = GRU_T(f_T).$$

[4] LSTM does not perform well

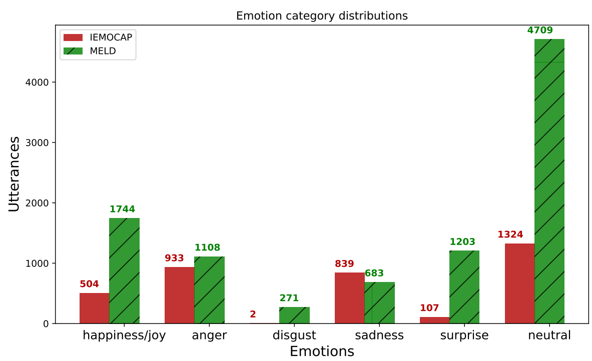

Figure 2: Comparison between the distribution of common emotions between training splits of IEMOCAP and MELD dataset.

4.3 Fusion

We then fuse F_A, F_T to a multimodal feature space. Since, the unimodal features may have different dimensions, we first equalize their dimensions and transform the features to dimensions D using a fully connected network.

$$g_A = \tanh(F_A W_A + b_A),$$
$$g_T = \tanh(F_T W_T + b_T),$$

where $W_A \in \mathbb{R}^{D_A\times D}$, $b_A \in \mathbb{R}^D$, $W_T \in \mathbb{R}^{D_T\times D}$, and $b_T \in \mathbb{R}^D$. We can represent the mapping for each dimension as

$$g_x = \begin{bmatrix} c^x_{11} & c^x_{21} & c^x_{31} & \cdots & c^x_{D1}\\ c^x_{12} & c^x_{22} & c^x_{32} & \cdots & c^x_{D2}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ c^x_{1N} & c^x_{2N} & c^x_{3N} & \cdots & c^x_{DN} \end{bmatrix},$$

where $x \in \{A, T\}$ and c^x_{lt} are scalars for all $l = 1, 2, \ldots, D$ and $t = 1, 2, \ldots, N$.

g_A, g_T were fused as follows -

$$f_{AT} = (f_{AT1}, f_{AT2}, \ldots, f_{AT(N)}),$$
$$\text{where } f_{ATt} = (i^{AT}_{1t}, i^{AT}_{2t}, \ldots, i^{AT}_{Dt}),$$

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

4.4 Classification and Training

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

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

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

As a regularization method, dropout between the GRU cell and dense layer is introduced to avoid overfitting.

Algorithm 1 Contextual Multomodal Emotion Classification Algorithm

procedure TRAINANDTESTMODEL(U, V) ▷ U = train set, V = test set
  Unimodal feature extraction:
  for i : [1, N] **do** ▷ extract baseline features
    $f_A^i \leftarrow AudioFeatures(u_i)$
    $f_T^i \leftarrow TextFeatures(u_i)$
  for m $\in \{A, T\}$ **do**
    $F_m = GRU_m(f_m)$
  Fusion:
  $g_A \leftarrow MapToSpace(F_A)$ ▷ dimensionality equalization
  $g_T \leftarrow MapToSpace(F_T)$
  $f_{AT} \leftarrow Fusion(g_A, g_T)$ ▷ bimodal fusion
  for i : [1, N] **do** ▷ softmax classification
    $\hat{y}^i = \underset{j}{\text{argmax}}(softmax(F_{AT}^i)[j])$
  $TestModel(V)$
procedure MAPTOSPACE(x_z) ▷ for modality z
  $g_z \leftarrow \tanh(W_z x_z + b_z)$
  return g_z
procedure FUSION(g_{z_1}, g_{z_2}) ▷ for modality z_1 and z_2, where $z_1 \neq z_2$
  for i : [1, D] **do**
    $f_{z_1 z_2}^i \leftarrow \tanh(w_i^{z_1 z_2} . [g_{z_1}^i, g_{z_2}^i]^\intercal + b_i^{z_1 z_2})$
  $f_{z_1 z_2} \leftarrow (f_{z_1 z_2}^1, f_{z_1 z_2}^2, \ldots, f_{z_1 z_2}^D)$
  return $f_{z_1 z_2}$
procedure TESTMODEL(V)
  Similarly to training phase, V is passed through the learnt models to get the features and classification outputs.

4.5 Baseline Results

We provide the baseline result in Table 9 and 10. As it can be seen from the tables that textual modality outperforms audio modality for both sentiment and emotion classification tasks. Surprisingly, multimodal and text baselines are very close. We believe this is due to the effect of a simple concatenation based fusion method. We also noted that many utterances in this dataset are relatively shorter for e.g., "Ok!", "Thanks" etc. which make the classification difficult. One solution could be the use of visual features with the textual features. We provide the video and audio files with the dataset which researchers should be able to utilize for better feature extraction method. The dataset is publicly available at `http://bit.ly/MELD-raw`.

5 Future Directions and Conclusions

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 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.

Modality	Sentiments			
	positive	negative	neutral	w-avg.
text	50.27	58.25	77.91	66.02
audio	6.33	13.25	66.12	39.72
text+audio	56.15	57.23	75.89	66.92

Table 9: Test-set F-score results of contextual biLSTM for sentiment classification in MELD. Note: *w-avg* denotes weighted-average.

5.1 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 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 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, 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.

Acknowledgments

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

Modality	Emotions							
	anger	disgust	fear	joy	neutral	sadness	surprise	w-avg.
text	43.22	0.23	0.56	52.73	75.37	21.47	49.47	56.72
audio	26.34	0.13	0.57	9.78	67.58	0.87	15.58	39.23
text+audio	43.12	0.34	0.45	52.47	77.37	11.58	51.32	57.34

Table 10: Test-set F-score results of contextual biLSTM for emotion classification in MELD. Note: *w-avg* denotes weighted-average.

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008a. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335–359.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008b. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

Chen, S.-Y.; Hsu, C.-C.; Kuo, C.-C.; Ku, L.-W.; et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

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

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 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, 2122–2132.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 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, 994–1003.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pérez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013. Utterance-level multimodal sentiment analysis. In *ACL (1)*, 973–982.

Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 873–883.

Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; and Pantic, M. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 449–456. ACM.

Teh, V., and Hinton, G. E. 2001. Rate-coded restricted boltzmann machines for face recognition. In Leen, T.; Dietterich, T.; and Tresp, V., eds., *Advances in neural information processing system*, volume 13, 908–914.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. 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, 2236–2246.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.