MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations

Mauajama Firdaus*, Hardik Chauhan, Asif Ekbal and Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
(mauajama.pcs16,hardik,asif,pb)@iitp.ac.in

Abstract

Emotion and sentiment classification in dialogues is a challenging task that has gained popularity in recent times. Humans tend to have multiple emotions with varying intensities while expressing their thoughts and feelings. Emotions in an utterance of dialogue can either be independent or dependent on the previous utterances, making the task complex and interesting. Multi-label emotion detection in conversations is a significant task that provides the ability to the system to understand the various emotions of the users interacting. On the other hand, sentiment analysis in dialogue or conversation helps in understanding the perspective of the user with respect to the ongoing conversation. Besides text, additional information in the form of audio and video assists in identifying the correct emotions with the appropriate intensity and sentiments in an utterance of a dialogue. Lately, quite a few datasets have been made available for emotion and sentiment classification in dialogues. Still, these datasets are imbalanced in representing different emotions and consist of only a single emotion. Hence, we present at first a large-scale balanced Multimodal Multi-label Emotion, Intensity, and Sentiment Dialogue dataset (MEISD) collected from different TV series that has textual, audio, and visual features, and then establish a baseline setup for further research.

1 Introduction

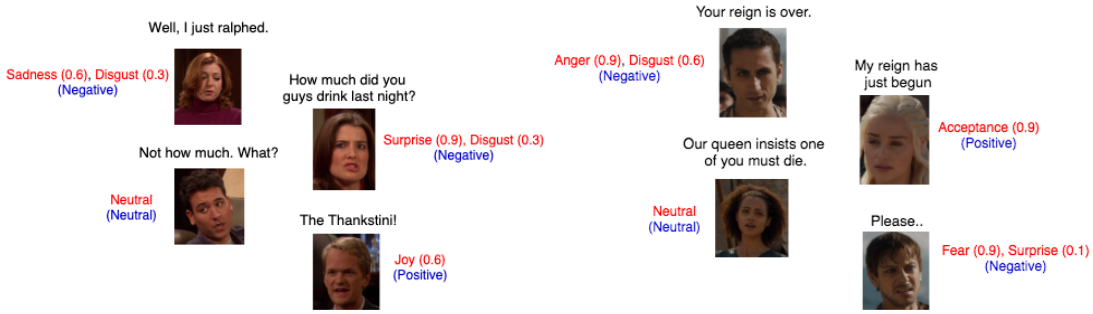
With the advancements in Artificial Intelligence (AI), the gap between Natural Language Processing (NLP) and Computer Vision (CV) has been bridged by extensive research in multi-modal information analysis. The ability to use different modalities such as text, audio and video for different tasks, such as emotion classification (Tripathi and Beigi, 2018; Hazarika et al., 2018a), sentiment analysis (Poria et al., 2017), dialogue generation (Yoshino et al., 2019; Das et al., 2017) have helped in building robust systems. The potential to understand correct emotion and sentiment in a conversation is crucial for developing strong human-machine interaction systems. Dialogue systems are of two types i.e., goal-oriented systems (Asri et al., 2017) or open chit-chat systems (Serban et al., 2017). In both these systems, understanding the user's emotions is crucial to maximizing the user experience and satisfaction. Nowadays, there is a huge demand for developing social agents capable of having real conversations with humans. With the rapid growth in technology, personal assistants in smartphones such as Amazon's Alexa, Apple's Siri, and Google's Home have become human companions. Hence, these applications need to understand the correct emotional state of the user to increase user contentment leading to user retention.

Emotions and sentiments are subjective qualities and are understood to share overlapping features; hence are frequently used interchangeably. This is mainly because both sentiment and emotion refer to experiences resulting from the combination of biological, cognitive, and social influences. Though both are considered to be the same, yet according to (Munezero et al., 2014), the sentiment is formed and retained for a longer duration, whereas emotions are like episodes that are shorter in length. Moreover, the sentiment is mostly target-centric, while emotions are not always directed to a target. Previously, sentiment and emotions have been tackled separately, although they are different but closely related.

Lately, emotion detection and sentiment analysis in multimodal systems using audio, video, and textual features have gained popularity. But both these tasks have not been explored in depth for conversations. The main reason for this is the unavailability of a large-scale multi-modal dialogue dataset labeled with emotions and sentiment to facilitate research in this direction. Also, identifying emotions and sentiments in conversations is a challenging task compared to tweets or sentences. This is mainly because the contextual information or past utterances may influence the emotions of the present utterance. Also, emotional state change among the speakers in a conversation makes it difficult to identify the emotions and sentiment of an utterance in a dialogue.

With the release of Multimodal EmotionLines Dataset (MELD), research in emotion and sentiment identification in conversations has gained immense attention. This dataset comprises the conversations taken from the Friends TV series labeled with sentiment and emotion using text, audio, and video information. The dataset provides multimodal information for classifying emotions and sentiments in dialogues. This dataset is made using a comedy TV series; it is unbalanced in its emotion distribution, making the dataset imbalanced. Human emotions are extremely complex; therefore, it is highly probable that they express multiple emotions in a single utterance. There is a huge possibility that multiple emotions expressed in an utterance are correlated. For example, the speaker may express the emotion "anger" and "disgust" often together than in isolation. Also, the intensity of the different emotions in a given utterance may vary. For example, the speaker, in some cases, express "anger" with higher intensity while "disgust" with lower intensity or vice-versa. The MELD dataset is labeled with a single emotion only, thereby not providing the complete emotional information in a given utterance.

For building robust emotion and sentiment classification systems, it is crucial to have a balanced dataset labeled with sentiment and multiple emotions along with their corresponding intensity to provide the complete affective information of a given utterance. Hence, in this work, we propose a large-scale balanced Multimodal Multi-label Emotion, Intensity, and Sentiment Dialogue (MEISD) dataset labeled with multiple emotions, intensity, and sentiments using textual, audio, and visual information, collected from 10 TV series belonging to different genres. Only textual information is not enough for understanding emotions, as emotion is also expressed through facial expressions, gestures, pitch, and tone. For example, the given utterance "Great, you are here" can exhibit different emotions, such as joy, anger, or surprise. Hence, it is difficult to identify the correct emotion using only the textual information. Hence, the sentiment label of these utterances is also ambiguous. It is essential to simultaneously focus on these utterances' audio and visual counterparts for identifying the correct emotions and sentiment label of these utterances. An example of a conversation from the MEISD dataset labeled with sentiment and multiple emotions, and their corresponding intensity is given in Figure 1. As it is evident from the given example, visual information provides additional knowledge for determining the correct emotions and sentiment labels. To the best of our knowledge, this is the first dialogue data labeled with multiple emotions, intensity, and sentiment for identifying emotions and sentiments in conversations and will hopefully promote further research in this area.



(a) Example 1 (b) Example 2

Figure 1: Examples from the MEISD dataset. Text in red represents the emotions with the corresponding intensity while text in blue represents the sentiment of the given utterance.

The major contributions of our present work are:

- We create a large-scale Multi-label Emotion, Intensity, and Sentiment Dialogue (MEISD) dataset for the task of multiple emotion, intensity, and sentiment classification in conversations.
- We provide some strong baselines for the proposed MEISD dataset for all the three tasks, *viz.* multi-label emotion classification, intensity prediction, and sentiment analysis on dialogues.

The rest of the paper is structured as follows. In Section 2, we present a brief survey of the related work. In Section 3, we describe the details of the dataset that we create. In Section 4, we explain the methodology. The experimental setup, along with the evaluation metrics, is reported in Section 5. In Section 6, we present the results along with the necessary analysis. Finally, we conclude in Section 7 with future work directions.

2 Related Work

Most of the early research on emotion classification and sentiment analysis was performed separately upon textual datasets mostly taken from twitter (Agarwal et al., 2011; Socher et al., 2013; Colnerič and Demsar, 2018; Ghosal et al., 2018; Chauhan et al., 2019). In (Chauhan et al., 2019), the authors proposed a RNN framework capable of learning inter-modal interaction among the different modalities using the auto-encoder mechanism. As emotion and sentiment are two very closely related tasks, in recent time there is a trend on modeling both sentiment and emotion of an utterance simultaneously (Akhtar et al., 2019a; Akhtar et al., 2019b; Kumar et al., 2019; Akhtar et al., 2020). In (Akhtar et al., 2020), the authors employed the concept of multi-task learning for multi-modal affect analysis and explored a contextual inter-modal attention framework that aimed in leveraging the association among the neighboring utterances and their multi-modal information. With the advancements in Artificial Intelligence (AI), emotion classification and sentiment analysis have become a significant task due to its importance in many downstream tasks, such as customer behavior modeling, response generation for conversational agents, multimodal interactions etc. Hence, to maximize user satisfaction and providing a better experience to the customer, it is important to understand the correct emotion and sentiment of the customer. Recently, multi-label emotion classification has been investigated for textual data in (Kim et al., 2018; He and Xia, 2018; Yu et al., 2018; Huang et al., 2019). Using multiple Convolution Neural Network (CNN) networks along with self-attention, the authors in (Kim et al., 2018) performed multi-label emotion classification on twitter data. Similarly, the authors in (Yu et al., 2018) improved the performance of multi-label emotion classification on twitter data by using transfer learning. Lately, sequence-to-sequence framework (Huang et al., 2019) has been employed for multi-label emotion classification. Our present work differs from these single and multi-label emotion and sentiment classification works as we tend to classifying emotions and sentiments on dialogue conversations that require contextual information of the previous utterances, thereby making the task more challenging and interesting.

Every human-machine interactions are grounded in conversations driven by emotions. Hence, identifying the emotion in dialogue is essential for building robust systems capable of such interactions. Recently, investigations on emotion detection in conversations has been in demand. The authors in (Chen et al., 2018) released a dataset taken from Friends TV series for detecting emotions in dialogues. Similarly, in (Yeh et al., 2019) an attention framework was designed for identifying emotions in spoken dialog systems. In (Hazarika et al., 2018b), memory networks were adopted to capture contextual information for emotion detection in conversations. To capture the contextual information in conversations, DialogueRNN (Majumder et al., 2019) employs three gated recurrent units (GRU) for effectively modeling the past utterances of the speaker and the listener in dyadic conversations for emotion detection.

As conversation itself is multimodal, people involved in conversations use various facial expressions, gestures and different pitch, tones to emote their feelings making the conversation dependent on the audio and visual aspect as well. Hence, quite a few multimodal datasets have been employed to identify emotion using audio and visual information as well. In (Hazarika et al., 2018a), the author proposed an interactive memory network that extracts multimodal features for emotion classification. IEMOCAP dataset (Tripathi and Beigi, 2018) has been used for emotion detection using a deep neural framework that uses the multimodal information at the final layer for emotion identification. Multimodal sentiment

analysis has also been investigated for correct classification of sentiments (Poria et al., 2017; Majumder et al., 2018). The authors in (Majumder et al., 2018) proposed a novel hierarchical feature fusion strategy for integrating different modalities, such as audio, video and text for identifying the sentiments. The authors in (Poria et al., 2019) extended the EmotionLines dataset by incorporating audio and visual modalities for correct identification of emotions and sentiments in conversations. The MELD dataset has been further used for building different neural frameworks for jointly identifying emotion and sentiment from conversations (Ghosal et al., 2019; Zhang et al., 2019b; Zhang et al., 2019a). As opposed to these existing works on multimodal emotion and sentiment classification on dialogue data, our present works provides a balanced multimodal multi-label emotion, intensity and sentiment dataset for the classification of multiple emotions and sentiment in the given utterance.

3 Multimodal Multi-label Emotion, Intensity and Sentiment Dialogue (MEISD) Dataset

We create the MEISD dataset[1] from the 10 famous TV shows belonging to different genres: (i). Comedy: Friends, The Big Bang Theory, How I Met Your Mother, The Office; (ii). Drama: House M.D., Grey's Anatomy, Castle and Game of Thrones, House of Cards, Breaking Bad. This dataset consists of conversations with utterances from multiple speakers making it a multi-party conversational dataset. The dataset contains dialogues mostly from all the episodes belonging to the different seasons of the TV series giving us a wide variation in dialogues. In total, we have 1000 dialogues from all the TV series in our dataset. Firstly, we obtain the start and end timestamps of every dialogue from the different episodes of the TV series. We extract all the subtitles and transcripts for every dialogue with their respective timestamps. Thereafter, we segment the dialogues into utterances following the heuristics similar to (Poria et al., 2019): (i). The timestamps of the utterances belonging to a dialogue should always be in the increasing order; (ii). The utterances in a particular dialogue should be from the same episode only. Utterances in the subtitles were sometimes grouped together under the same timestamp in the subtitle files. Hence, we use the transcription alignment tool Gentle[2] for extracting the accurate timestamp information of every utterance as it automatically aligns the text with the audio by obtaining the word-level timestamp information from the audio file. After extracting the corresponding timestamps of every utterance in a dialogue, we then obtain their audio and visual clips from the source episodes. After getting the audio and visual clips of every utterance, we extract the audio and visual files from these clips. The audio files are then formatted as 16-bit PCM WAV files for further processing. The video files were used to extract 2048D pooled features using the last convolution block of ResNet101. Our final MEISD dataset comprises of textual, visual and audio features that bring the three important modalities together for effective multi-label emotion, intensity and sentiment analysis.

3.1 Annotation

The utterances in every dialogue of the MEISD dataset is annotated with the appropriate emotion category and their corresponding intensity. For annotating the dataset, we consider Ekman's (Ekman, 1992) six universal emotions, namely Joy, Sadness, Anger, Fear, Surprise, and Disgust as emotion labels for the utterances in a dialogue. The emotion annotation list has been extended to incorporate two more labels, namely *Acceptance* and *neutral*. The "acceptance" emotion has been taken from the Plutchik's (Plutchik, 1980) wheel of emotions for utterances in a dialogue expressing this emotion while the "neutral" label is designated to utterances having no-emotion. Every emotion label is accompanied with an intensity value ranging from 1-3, with 1 indicating the lower intensity and 3 the highest. Every utterance in a given dialogue is labeled with sentiment labels (i.e. positive, negative and neutral) as well.

For annotating the utterances in our dataset, we employ four graduate students highly proficient in English comprehension. The guidelines for annotation along with some examples were explained to the annotators before starting the annotation process. As we create a multimodal dataset, hence the annotation of the dataset was also done in a similar manner. The data was annotated by not just looking at the transcripts (textual information) but also focusing on the audio and visual clips of the corresponding

[1]The MEISD dataset is available at `https://www.iitp.ac.in/~ai-nlp-ml/resources.html\#meisd`

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

Categories		MEISD		
		Train	*Valid*	*Test*
Emotion	*Anger*	2145	294	577
	Disgust	1723	301	471
	Joy	2589	331	533
	Surprise	2216	315	587
	Acceptance	1562	214	439
	Sadness	1433	268	460
	Fear	1067	170	451
	Neutral	1417	208	429
Sentiment	*Positive*	4968	800	1489
	Negative	5717	983	1579
	Neutral	3417	318	929

Table 1: Emotion and Sentiment distribution

Statistics	Train	Valid	Test
# Modalities	(t,a,v)	(t,a,v)	(t,a,v)
# Dialogues	702	93	205
# Utterances	14040	1860	4100
# Speakers	2418	632	1022
Avg. Utterance length	12	10.5	11.7
Avg. # of utterances per dialogue	20.2	19.8	20.1
Avg. # of emotions per dialogue	4.5	4.2	4.7
Avg. # of emotions per utterance	2	2	2
# of unique words	25781	7189	17458
Avg. duration of an utterance	4s	3.58s	4s

Table 2: Dataset statistics. Here, (t,a,v) = (text,audio,video)

utterance. Hence, for every utterance, the annotators were asked to watch the video clip and listen to the audio files along with the text for annotating the utterance with the appropriate emotion and sentiment labels. The annotators were also given the contextual information (text, audio and video) for a given utterance for reference so that they are able to provide correct emotion and sentiment labels.

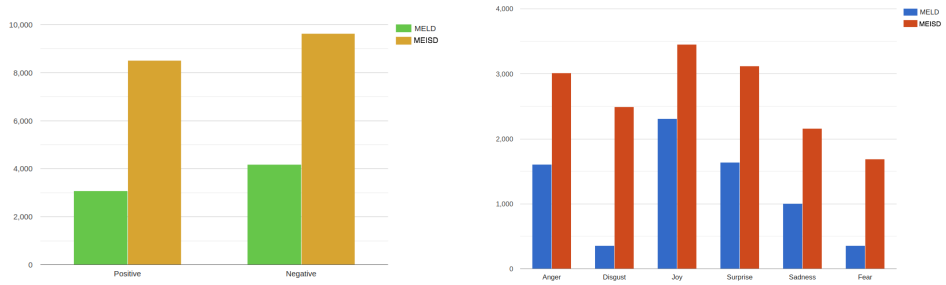
Majority voting scheme was used for selecting the final emotions or sentiment label for each utterance. We achieve an overall Fleiss' (Fleiss, 1971) kappa score of 0.67 for the emotions, 0.72 for intensity and 0.75 for sentiment which can be considered as reliable. The use of audio and visual modalities for annotation has helped in achieving the correct emotion labels with the corresponding intensity for every utterance of the dialogue. The utterances for which the annotators could not reach an agreement on the emotions, intensity or sentiment labels were removed from the dataset to avoid any discrepancies in the data. In Table 1, we show the overall emotion and sentiment distribution of our dataset.

Utterance	Emotion	Sentiment
And live forever as a machine!	Disgust	Positive
Look at you, all jealous.	Joy	Negative
Brain tumors at her age are highly unlikely	Sadness	Positive
Your political consultants have written you a nice story	Disgust	Positive
I bet it was one of her backstabbing rivals	Acceptance	Negative

Table 3: Examples from the MEISD dataset showing contrasting emotion and sentiment labels for a given utterance

As already mentioned, we annotate our dataset with eight emotion labels, i.e. anger, disgust, fear, joy, acceptance, neutral, sadness, and surprise with an intensity range from 1-3 and three sentiment labels i.e. positive, negative and neutral. From the emotion distribution given in Figure 2b, it is evident that the emotion labels are balanced in comparison to the MELD dataset as we have extracted the dialogues from different TV series, hence providing diversity in dialogues. The sentiment labels of the utterances were also annotated along with emotion. The sentiment distribution of both the datasets is given in Figure 2a.

As already mentioned the authors in (Poria et al., 2019) labeled the utterances with single emotion while losing the information of other possible emotions present in the given utterance. Also, the authors in (Poria et al., 2019) labeled every utterance with sentiments based on their emotion labels. Positive sentiment label was given to the utterances having joy as the emotion label and negative sentiment was labeled to the utterances having anger, disgust, sadness, fear as emotion labels. While they only annotated the surprise emotion label with sentiments having positive and negative sentiment labels as this emotion is considered to fall on either of the sentiment labels. Hence, we take care of the fact that the sentiment is annotated independently without being biased on the emotion label. From the example, given in Table 3, we can see that the sentiment label and emotions are independent at times, whereas a positive sentiment label can be given to negative emotion and vice versa. Hence, in preparing our MEISD dataset, we have taken care of these details as sentiment or emotion is dependent on the contextual information and the speaker of the utterance.



(a) Sentiment Distribution

(b) Emotion Distribution

Figure 2: Sentiment and Emotion distribution of MELD vs. our Proposed MEISD dataset

In Table 2, we provide the important statistics of the MEISD dataset. The average duration of an utterance in our dataset is approximately 4 seconds. The average length of an utterance in a dialogue across the training, validation and test sets are almost the same. The average dialogue length comprises of 20 utterances and it is the same across the training, validation and test splits. Every dialogue on an average consists of five emotions while the average number of emotions in a given utterance is 2. The presence of multiple speakers and the emotion shift of a speaker makes the task of emotion and sentiment analysis very interesting as well as challenging. In Figure 3, we show the emotion shift of a speaker as the dialogue grows.

Figure 3: A dialogue from the MEISD dataset showcasing the emotion shift as the conversation grows. The text in blue represents the sentiment label while the text in red represents the emotion label of every utterance.

3.2 Comparison with Related Datasets

The available datasets for multimodal emotion detection and sentiment classification are non-conversational. The examples of such datasets are MOUD (Pérez-Rosas et al., 2013), MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018) that have been deeply investigated by the researchers for both the tasks. Two dyadic conversational datasets, IEMOCAP (Busso et al., 2008) and SEMAINE (McKeown et al., 2011) have gained popularity for encouraging research on emotion detection for conversations. Recently, MELD (Poria et al., 2019) dataset was released to inspire research on multiparty conversations using information from different modalities.

IEMOCAP Dataset: The IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) dataset (Busso et al., 2008) comprises of videos of dyadic interactions between pairs of 10 speakers across a duration of 10 hours having different dialog situations. The utterances are extracted by segmenting the videos and then labeling each utterance with fine-grained emotion labels, such as anger, excitement, happiness, frustration, neutral, and sadness. The dataset also gives continuous attributes in the form of valence, activation, and dominance for facilitating better emotion detection of the utterances. Our MEISD dataset differs majorly from this dataset as ours is labeled with multiple emotions, intensity and sentiment categories to jointly perform both sentiment and emotion tasks.

SEMAINE Dataset: The SEMAINE dataset (McKeown et al., 2011) is an audiovisual database designed to engage a person in a continuous and emotional conversation. The conversations in the dataset comprise interactions concerning a human and an operator (where it can be either a person or a person simulating a machine). In total, there are 150 participants in the dataset, having 959 conversations, where each conversation having a duration of about 5 minutes. This dataset is different from our proposed dataset as we provide multiparty conversations labeled with both sentiment and emotion labels.

MELD Dataset: The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) comprises of multiparty conversations taken from the Friends TV series. The dataset has been annotated with 7 emotion labels, namely anger, fear, disgust, surprise, neutral, sadness, and joy. The dataset has also been annotated with three sentiment labels i.e., positive, negative and neutral. The dataset comprises of 13,000 utterances having textual, audio and visual information, hence facilitating multimodal research for emotion and sentiment in multiparty conversations.

Our proposed dataset, though having multiparty conversations with multimodal information is different from the MELD dataset. The dataset that we present here is larger compared to MELD. The major difference being that we provide multi-label emotion information with the corresponding intensity for the utterances in a dialogue. Our emotion labels are balanced in comparison to the MELD dataset, since we have taken conversations from different TV series. By using different TV series belonging to different genres, we provide diversity in our dataset. Hence, every emotion is depicted by various characters that bring diverseness in the way a particular emotion is expressed making the task exciting as well as challenging. Comparisons between the existing datasets and our proposed MEISD dataset are given in Table 4.

Dataset	**Type**	**No. of Dialogues**			**No. of Utterances**		
		Train	*Valid*	*Test*	*Train*	*Valid*	*Test*
SEMAINE	acted	58		22	4386		1430
IEMOCAP	acted	120		31	5810		1623
MELD	acted	1039	114	280	9989	1109	2610
MEISD	acted	702	93	205	14040	1860	4100

Table 4: Comparison of different multimodal conversational datasets and our proposed MEISD dataset

4 Experiments

The extraction of features along with the details of the baseline models to evaluate our proposed MEISD dataset is described in this section. We also discuss the metrics used to evaluate the models on the proposed dataset.

4.1 Feature Extraction

Textual Features: For textual features, we take the pre-trained 300-dimensional GloVe embeddings (Pennington et al., 2014) of every word as features.

Audio Features: We encode audio tracks with the pre-trained VGGish network (Hershey et al., 2017), which is trained on Audioset (Gemmeke et al., 2017) consisting of 100 million YouTube videos. It has been shown to improve the audio emotion and sentiment classification. We extract audio features of dimension 128 from the last fully connected layer.

Visual Features: Due to computational cost, we only consider the middle frame of the video to extract visual feature V_k. We use 2048-dimension pooled features from the last block of Resnet-101 (He et al., 2016) pre-trained on Imagenet (Russakovsky et al., 2014) for visual features.

The bimodal or the multimodal features are obtained by concatenating the respective audio, visual and textual features as needed in the model.

4.2 Baseline Models

In order to provide strong baselines for our MEISD dataset, we perform several experiments with different baselines. We extend the existing baselines for multi-label emotion and intensity prediction. We

model multi-label emotion, sentiment as the classification; and intensity prediction as the regression task, respectively. All the implementations are done using the PyTorch [3] framework. Based on the validation set, we set the threshold value of 0.2 for the classification of multiple emotions in a given utterance. For all the baselines, in the final output layer we apply softmax activation function for emotion and sentiment classification while we apply sigmoid activation function for intensity prediction.

text-CNN: In this approach, we only use the textual information for identifying the emotion and sentiment of every utterance in a dialogue. In this framework, we use the word embeddings of the utterances as input to the convolutional neural network (CNN) (Kim, 2014) for obtaining the sentence representation. In this model, we do not use the contextual information or the additional information from the different modalities for identifying the emotion or sentiment of an utterance.

bcLSTM: This baseline employing bi-directional RNN for capturing the contextual information was proposed by (Poria et al., 2017). It employs a two-step hierarchical mechanism that captures the uni-modal context first followed by the bi-modal context features. In this methodology, we incorporate the provision of capturing information from all the three modalities. A CNN-LSTM approach is used for unimodal text to extract the textual features using the Glove embeddings as input to the model. For audio representations, we employ a LSTM with every audio feature vector as input to the model. Similarly, for video representations, we employ a LSTM model giving the visual feature vector as the input. Finally, the representations from the unimodal are fed as input to the multimodal framework for identifying the corresponding emotion, sentiment and intensity of the utterance.

DialogueRNN: This baseline proposed by (Majumder et al., 2019) is one of the current state-of-the-art approaches for modelling emotions and sentiments in conversations. It is a powerful baseline for modeling context with effective mechanisms by tracking individual speaker states throughout the dialogue for correct emotion and sentiment classification. Since DialogueRNN can handle multi-party interactions, hence it can be applied directly to our proposed MEISD dataset. It utilizes three levels of gated recurrent units (GRU) to model conversational context for correctly identifying the emotions, intensity and sentiments in a dialogue.

DialogueRNN + BERT: We propose a stronger baseline built upon the DialogueRNN for correct classification of emotion and sentiment, and for intensity prediction. We are able to improve the performance of DialogueRNN by using BERT(Devlin et al., 2018) embedding instead of Glove embedding to represent the textual features.

4.3 Evaluation Metrics

For multi-label emotion classification, we use the automatic metrics as mentioned below following the works of (Huang et al., 2019; Yang et al., 2018; Mohammad et al., 2018): Jaccard Index (Rogers and Tanimoto, 1960), Hamming Loss (Schapire and Singer, 1999) and Micro-averaged F1-score (Manning et al., 2008). For sentiment analysis we report Micro-averaged F1-score while for intensity prediction we report Pearson correlation co-efficient (Mohammad and Bravo-Marquez, 2017) in a similar manner as (Akhtar et al., 2019b).

5 Result and Discussion

In this section, we provide the results for all the three tasks, i.e. multi-label emotion classification, intensity prediction and sentiment analysis on our proposed MEISD dataset. In Table 5, we provide the results of all the three tasks for all the different baselines.From the results, it is evident that we achieve a weighted overall F1 score of 62.29% using our proposed baseline which has been built upon the DialogueRNN. We have used BERT representations as the textual features which help in improving the performance of the model by increasing the F1 scores in case of multi-label emotion classification. In case of Jaccard index which is equivalent to multi-label accuracy, we see an improvement in the proposed baseline with an accuracy of 53.7%. Lower hamming loss in the proposed baseline indicates the better performance of the model for the given task.

[3] https://pytorch.org/

From the table, we can also infer that using solely the audio and video features of every utterance decreases the performance of the model in identifying the correct emotions. The major information about the emotions is achieved from the textual features itself, hence the performance of the models using only textual features is far better than the models having only audio and video features as input. While using all the features, they together boost the performance of the model. Hence, it can be concluded that the audio and visual counterparts of an utterance assist in identifying the correct emotions of a particular utterance. Since in our final baseline model we only enhance the performance by using better textual representation, hence the performance on audio and visual are similar to the DialogueRNN baseline. For the intensity prediction task, we report the Pearson correlation co-efficient as a metric and from the table it is visible that the final proposed baseline yields the highest score of 0.588 using information from all the three modalities.

Simultaneously, in Table 5 we present the results of sentiment classification on our dataset for the several baselines as mentioned in the previous section. Overall, we achieve F1 score of 69.25% from our DialogueRNN + BERT based baseline model. Even in the case of sentiment, we see that BERT helps in improving the overall performance of the individual sentiment labels, thereby enhancing the F1 score of the model. As almost all the sentiment labels are in equal proportion, hence the performance of each label is almost the same with respect to each other.

Models	**Modality**			**Multi-label Emotion Classification**			**Intensity Prediction**	**Sentiment Analysis**
	T	*A*	*V*	*JI*	*HL*	*M-F1*	*P-Corr*	*M-F1*
text-CNN	√	-	-	0.415	0.168	54.18	0.392	62.89
bc-LSTM	√	-	-	0.468	0.157	57.05	0.476	64.34
	-	√	-	0.342	0.213	41.17	0.311	38.85
	-	-	√	0.311	0.256	39.45	0.293	21.53
	√	√	√	0.495	0.145	59.32	0.481	65.21
DialogueRNN	√	-	-	0.471	0.151	58.73	0.485	65.59
	-	√	-	0.349	0.207	41.52	0.318	40.15
	-	-	√	0.321	0.243	40.87	0.305	22.33
	√	√	√	0.519	0.141	60.57	0.513	65.87
DialogueRNN + BERT	√	-	-	0.520	0.140	60.93	0.524	68.78
	-	√	-	0.351	0.205	41.52	0.337	40.15
	-	-	√	0.322	0.241	40.87	0.319	22.33
	√	√	√	0.537	0.136	62.29	0.588	69.25

Table 5: Results of different models on MEISD dataset for multi-label emotion classification, intensity prediction and sentiment analysis. Here, T: Text, A: Audio, V: Visual features; JI: Jaccard Index; HL: Hamming Loss; M-F1: Macro-averaged F1 score; P-Corr: Pearson Correlation

6 Conclusion and Future Work

In this paper, we have introduced a large-scale multimodal multiparty conversational dataset, MEISD for multi-label emotion classification, intensity prediction and sentiment analysis in conversations. The detailed description of the dataset along with the entire process for building the dataset has been discussed in the paper. MEISD dataset is a multimodal dataset that has textual, audio and visual features for every utterance of dialogue taken from 10 different TV series belonging to the different genres, thereby providing a large diversity to the MEISD dataset. Hence, this dataset provides diversity with respect to utterances, scene information, characters and emotional expressions, and hence offer a wide variety in dialogues and make the task all the more challenging. We have evaluated our proposed MEISD datasets and reported the results using strong baselines for all the three tasks of emotion recognition, intensity prediction and sentiment classification. We believe that this dataset can be employed in the future for multi-label emotion, intensity and sentiment detection in conversations.

In the future, this dataset can be employed for multi-task learning of all the three tasks simultaneously in dialogues. As all the three tasks are closely related, hence through multi-task learning the performance of the tasks might improve due to the shared information. This dataset can also be used for building emotional and sentimental conversational agents. Furthermore, the multimodality aspect of the dataset can be investigated deeply employing different fusion techniques for achieving better multi-modal interactions that can help in the tasks. Also, research on novel frameworks for capturing the contextual information

for better classification of all the tasks can be investigated in the future.

Acknowledgement

Authors duly acknowledge the support from the Project titled "Sevak-An Intelligent Indian Language Chatbot", Sponsored by SERB, Govt. of India (IMP/2018/002072). Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June. Association for Computational Linguistics.

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019a. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 370–379.

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2019b. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*.

Md Shad Akhtar, Dushyant Singh Chauhan, and Asif Ekbal. 2020. A deep multi-task contextual attention framework for multi-modal affect analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(3):1–27.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 207–219.

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.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5651–5661.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Niko Colnerič and Janez Demsar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11(3):433–446.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3454–3466.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part I*, pages 250–259. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE.

Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Seq2emo for multi-label emotion classification based on latent variable chains transformation. *arXiv preprint arXiv:1911.02147*.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. 2018. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 141–145.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. 2019. Emotion helps sentiment: A multitask model for sentiment and emotion analysis. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pages 1–8. IEEE.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, volume 33, pages 6818–6825.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 34–49.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 1–17.

Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on Affective Computing*, 5(2):101–111.

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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 973–982.

R Plutchik. 1980. Plutchik's wheel of emotions.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

David J Rogers and Taffee T Tanimoto. 1960. A computer program for classifying plants. *Science*, 132(3434):1115–1118.

Olga Russakovsky, Jun Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

Robert E Schapire and Yoram Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336.

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 *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642.

Samarth Tripathi and Homayoon Beigi. 2018. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3915–3926.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6685–6689. IEEE.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

Jianfei Yu, Luis Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1097–1102.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. 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. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2236–2246.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019a. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421. AAAI Press.

Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. 2019b. Quantum-inspired interactive networks for conversational sentiment analysis. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5436–5442.