More the Merrier: Towards Multi-Emotion and Intensity Controllable Response Generation

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

The focus on conversational systems has recently shifted towards creating engaging agents by embedding emotions into them. Human emotions are highly complex as humans can express multiple emotions with varying intensity in a single utterance, whereas the conversational agents currently convey only one emotion in their responses. To infuse human-like behaviour in the agents, we introduce the task of multi-emotion controllable response generation with the ability to express different emotions with varying levels of intensity in an open-domain dialogue system. We introduce a Multiple Emotion Intensity aware Multi-party Dialogue (MEIMD) dataset having 34k conversations taken from 8 different TV Series. We propose a framework with Multiple Emotion with Intensity-based Dialogue Generation (MEI-DG). The system employs two novel mechanisms: (i) determining trade-off between the emotion and generic words, while focusing on the intensity of the desired emotions; and (ii) computing the amount of emotion left to be expressed, thereby regulating the generation. Detailed evaluation shows that our proposed approach attains superior performance compared to the baseline models.

Introduction

Conversational agents like Apple's Siri, Microsoft's Cortana not only assist humans in completing tasks, but also behave as companion. It is becoming increasingly necessary to endow conversational agents with the ability to perceive and express emotions. These agents enhance user satisfaction (Prendinger, Mori, and Ishizuka 2005), while reducing breakdowns in conversations (Martinovski and Traum 2003) and providing for user retention. Hence, dialogue systems capable of generating replies, while considering the user's emotional state, is one of the most welcome advancements in Artificial Intelligence (AI).

Previously, researchers have focused on classifying user emotions (Poria et al. 2019; Chauhan et al. 2019) in conversations. For building an intelligent agent, understanding per se of emotion is insufficient. Hence several works (Song et al. 2019; Colombo et al. 2019) have concentrated on inducing emotion into the dialogue system. Most of these existing research have focused on generating emotionally aware (Rashkin et al. 2019; Lin et al. 2019) and emotionally controlled responses (Zhou et al. 2018; Huang et al. 2018; Firdaus et al. 2020c) expressing a single emotion or a particular intensity of the emotion (Ghosh et al. 2017).

Though many systems like the ones mentioned above can express a particular emotion, they lack completeness. Humans routinely express multiple feelings in their day-to-day conversations. Recent research has focused on identifying multiple emotions in user utterances (Yu et al. 2018; Huang, Trabelsi, and Zaïane 2020; Firdaus et al. 2020b). In addition to the problem of multiplicity, one faces the challenge of intensity. The intensity of emotion varies in dialogues. Existing conversational agents are inadequate in generating responses like "It's amazing, I am thrilled you got promoted" as shown in Table 1, since such responses contain multiple emotions, and the agents are can generate one emotion only From this example, it is also evident that intensity is an important dimension in response generation, as the intensity level varies in utterances. Even if the complete set of emotions is provided, the agent will have difficulty generating the response mentioned above as one emotion has a higher intensity than the other.

This perspective motivates our work. Conversational agents with the ability to express multiple emotions along with appropriate intensity provide enriched affective outlook. To mimic this human-like behaviour in conversational agents, we propose the task of generating responses conditioned on different emotions along with intensity.

Due to the unavailability of multi-emotion intensity labeled data for our proposed task, we create a large-scale dialogue dataset, MEIMD from 8 English TV series having 34k conversations that have been labeled with multiple emotions and their intensities. Examples of multi-label emotion conversations from our dataset are shown in Table 1. To solve the task of emotional text generation conditioned on multiple emotions with intensity, we propose a novel neural architecture- Multiple Emotion with Intensity-based Dialogue Generation (MEI-DG) framework- having implicit and explicit memory. Explicit memory governs the correct choice of words at every step for expressing the specified emotions. Implicit memory focuses on expressing the desired emotions by utilizing the VAD lexicon (Mohammad 2018) of every emotion. The implicit memory also balances emotion with grammatical correctness dynamically, for better generation.

	Conversations	Emotions
1	It's amazing, I am thrilled you got promoted I have loads of work now and am afraid to complete it. Stop sulking, I am sure you will manage it.	Surprise (0.3), Joy (0.9) Disgust(0.3), Fear(0.6) Anger(0.3), Acceptance(0.6)
2	I am sorry this could be an infection or cancer. I am afraid but I know you could help me.	Sadness(0.6), Fear(0.3) Acceptance(0.3), fear(0.6)

Table 1: Multi-Emotion and Intensity labeled conversations from MEIMD dataset

Contributions of our current work are: (i) to the best of our knowledge, we are the first to propose the task of multiple emotion and intensity controlled dialogue generation; (ii) we create a large-scale Multiple Emotion and Intensity aware Multi-party Dialogue (MEIMD) dataset; (iii) we design a neural architecture- MEI-DG- employing two novel memory-based mechanisms, *viz.*, implicit memory and explicit memory to ensure incorporation of multiple emotions with their corresponding intensity in the responses; (iv) empirical analysis shows that our proposed MEI-DG framework outperforms the baseline models and generates emotionally correct, rich responses.

Related Work

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 (Yeh, Lin, and Lee 2019; Chauhan et al. 2019; Poria et al. 2019) has been an important research direction. Currently, multi-label emotion classification has been investigated in (Kim, Lee, and Jung 2018; Yu et al. 2018; Huang, Trabelsi, and Zaïane 2020; Firdaus et al. 2020b) for understanding different emotions present in an user utterance. Inspired by these current works on emotion classification has led to emotional response generation research for building intelligent conversational agents.

Lately, emotional text generation has gained immense popularity (Huang et al. 2018; Li and Sun 2018; Lin et al. 2019; Li et al. 2017; Ghosh et al. 2017; Rashkin et al. 2019). In (Zhou et al. 2018), an emotional chatting machine (ECM) was built based on seq2seq framework for generating emotional responses. Recently, a lexicon-based attention framework was employed to generate responses with a specific emotion (Song et al. 2019). Emotional embedding, affective sampling and regularizer were employed to generate the affect driven dialogues in (Colombo et al. 2019). The authors employed curriculum dual learning (Shen and Feng 2020) for emotion controllable response generation. In (Asghar et al. 2018; Lubis et al. 2018; Zhong, Wang, and Miao 2019; Li et al. 2020), an end-to-end neural framework has been proposed that captures the emotional state of the user for generating empathetic responses. Our present research differs from these as we propose and address a novel task of generating responses with multiple emotions and intensity. To the best of our knowledge, this is the very first attempt to provide a benchmark setup for multi-emotion and intensity aware response generation in a dialogue setting.

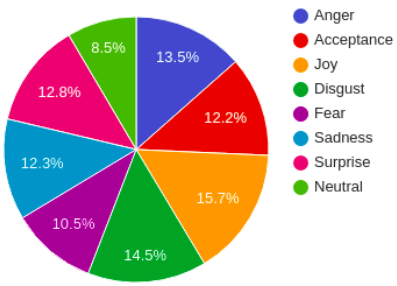

Figure 1: Emotion distribution of the MEIMD dataset

Multiple Emotion and Intensity aware Multi-party Dialogue (MEIMD) Dataset

The existing emotion labeled dialogue datasets (Rashkin et al. 2019; Li et al. 2017; Winata et al. 2019) have been marked with single emotions only without considering the intensity level of the emotion present in the utterance. For building a multi-emotion and intensity controlled framework, we create a large-scale multi-party dataset (MEIMD), one of the contributions of our current work. The MEIMD is a crowd-sourced dataset where every utterance in a given dialogue has been labeled with the corresponding emotions and intensity values to facilitate multi-emotion and intensity controlled response generation. The complete statistics of the MEIMD dataset are given in Table 2.

Data Collection: Previously Movie subtitles (Huang et al. 2018; Colombo et al. 2019; Asghar et al. 2018; Lubis et al. 2018; Moghe et al. 2018) and TV shows (Poria et al. 2019; Firdaus et al. 2020a) have been used for creating dialogue datasets for different Natural Language Processing (NLP) tasks. Inspired by these works, we create our MEIMD dataset by considering different TV shows having diverse conversations for building robust systems. For data collection, we consider 8 famous TV shows belonging to the different genres: (i). Drama: Breaking Bad, Castle, Game of Thrones, Grey's Anatomy, and House M.D.; (ii). Comedy: Friends, How I Met Your Mother and The Big Bang Theory. In total, there are 507 episodes, spanning 456 hours. First, we extract all the subtitles and transcripts for every episode. Then we segment the episodes into scenes which were further divided into short clips representing a dialogue. The dialogues were created following the heuristics in which the time-stamps of the conversation's utterances had to be in increasing order, and every statement in a dialogue should belong to the same scene of an episode by using the transcripts of every episode.

Data Preparation: To annotate the MEIMD dataset with multiple emotions and their corresponding intensity, we employ crowd-workers that label every utterance with the provided set of emotion labels and the intensity range. We consider 7 emotion labels *(anger, acceptance, disgust, fear, joy, sadness, surprise)* with intensity values ranging from *0-3*[1]

[1]Here, 0 means absence of emotion while 3 is the highest value of intensity expressing the maximum amount of that particular emotion.

Show	Genre	# Seasons	# Episodes	# Dialogues	# Utterances	# Main Speakers	Avg. Turns per Dialogue	Avg. Utterance Length	# of Emotions per Dialogue	Avg. emotions per Utterance
Breaking Bad	Drama	5	62	1659	32653	11	20.16	14.2	3.5	2
Castle	Drama	5	105	5172	102394	9	21.11	13.8	4.2	2
Friends	Comedy	10	236	4228	82353	6	23.40	10.6	5.5	2
Game of Thrones	Drama	7	67	2263	47471	30	22.50	13.7	3.8	2
Grey's Anatomy	Drama	6	126	4428	86104	15	22.17	14.5	4.1	2
House M.D.	Drama	8	177	6476	126780	12	21.43	13.6	3.3	2
How I Met Your Mother	Comedy	9	208	4968	96314	6	22.33	12.8	5.4	2
The Big Bang Theory	Comedy	10	207	5410	86913	7	21.98	12.5	5.3	2
Total	**-**	**60**	**1188**	**34604**	**660982**	**96**	**21.88**	**13.21**	**4.4**	**2**

Table 2: Dataset statistics of the proposed MEIMD dataset for every TV show

along with the *"Neutral"* label showcasing no emotion to annotate our MEIMD dataset. These emotion labels are chosen from Ekman's six basic emotions (Ekman et al. 1987) and Plutchik's wheel of emotions (Plutchik 2001). For labeling the utterances, the workers were asked to follow the instructions and guidelines provided for annotation. Some of the significant guidelines for annotation were as follows: (i). Every utterance of a given dialogue was to be marked with the provided emotion labels; (ii). In addition, the workers were asked to label as many emotions present in a given utterance capturing even the subtle emotion present. For cases where we found different annotations in emotions for a particular utterance, we remove them from the dataset, and we also drop the entire dialog to maintain coherence among the utterances. A majority voting scheme was used for selecting the final emotions for every utterance. We observed a multi-rater Kappa (McHugh 2012) agreement ratio of approximately 69% for the emotion, which can be considered reliable. (iii). After final emotion annotations the workers were also asked to annotate the intensity value of the given emotion from the specified range. Similarly, majority voting with the kappa ratio of 57% was observed for intensity which can be considered as reliable. The emotion distribution of the entire MEIMD dataset is provided in Figure 1. In Figure 1, we see that the MEIMD dataset is balanced, having almost equal representation of all the different emotions. In Figure 2, we provide the distribution of multiple emotions in the MEIMD dataset. It is evident that there are many instances of utterances having more than one emotion that motivates us to undertake our proposed task.

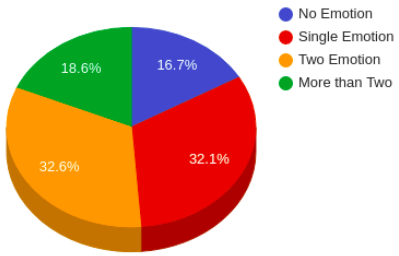

Figure 2: Number of Emotion distribution in the dataset

Methodology

For the given task, we assume that the emotion categories and the corresponding intensity will be provided to generate the response. As emotions are subjective and majorly dependent upon the respondent, there can be multiple suitable responses possible for a given input. Due to the intricacy in human emotions, we propose multiple emotions with intensity-dialogue generation (MEI-DG) network as our generation framework, as shown in Figure 3.

Problem Formalization: Given a dialogue having k turns $D = (U_1, U_2, ...U_k)$ as an input where each turn comprises of $U_k = (w_{k,1}, w_{k,2}, ..., w_{k,n})$ words along with the emotion categories $e_1, e_2...e_{n"}$ and intensity $I_1, I_2...I_{n"}$ the task is to generate the next response $Y = (y_1, y_2, ..., y_{n'})$ which is in accordance to the specified emotions and intensity and also coherent with the conversation, where $w_{k,n} \in V$ and $y_{n'} \in V$ are words in the conversational history and the generated response, respectively. Here, n and n' denote the given input utterance and response length, respectively, while $n"$ is the total number of desired emotions with their intensity. We design a vocabulary consisting of emotional words E_v and generic words G_v such that $V = E_v \bigcup G_v$ and $E_v \bigcap G_v = \phi$. The emotional vocab E_v is further split into multiple subsets E_v^i of words belonging to a particular emotion category i. For simplicity, we consider only two emotions and corresponding intensity. Hence $n"$= 2.

Multiple Emotion with Intensity Dialogue Generation (MEI-DG) Framework: It is built upon the Hierarchical Encoder-Decoder (HRED) architecture (Serban et al. 2015) to incorporate multiple emotions with varying intensity in the generated responses. In our proposed framework, we use the RNN (LSTM) network for encoding the context and utterance information.

Utterance Encoder: Given an utterance U_k, a bidirectional LSTM (BiLSTM) is employed to encode each word $w_{k,i}$, $i \in (1, ..., n)$ represented by d-dimensional embeddings. We concatenate the last hidden representation from both unidirectional LSTMs to form the utterance's final hidden representation.

$$s_{u,k,i} = BiLSTM_u(w_{k,i}, s_{u,k,i-1}) \quad (1)$$

Context-level Encoder: The information of the utterance encoder for every dialog turn serves as input to the context encoder. We use uni-directional LSTM to model dialog history. The final hidden state of the context LSTM serves as the initial state of the decoder LSTM.

$$s_{c,i} = LSTM_c(s_{u,i,n}, s_{c,i-1}) \quad (2)$$

Decoder: In the decoding stage, we employ unidirectional LSTM that generates words sequentially conditioned on the context vector c_t, implicit memory states of previous time step M_{t-1}, the desired emotion embeddings,

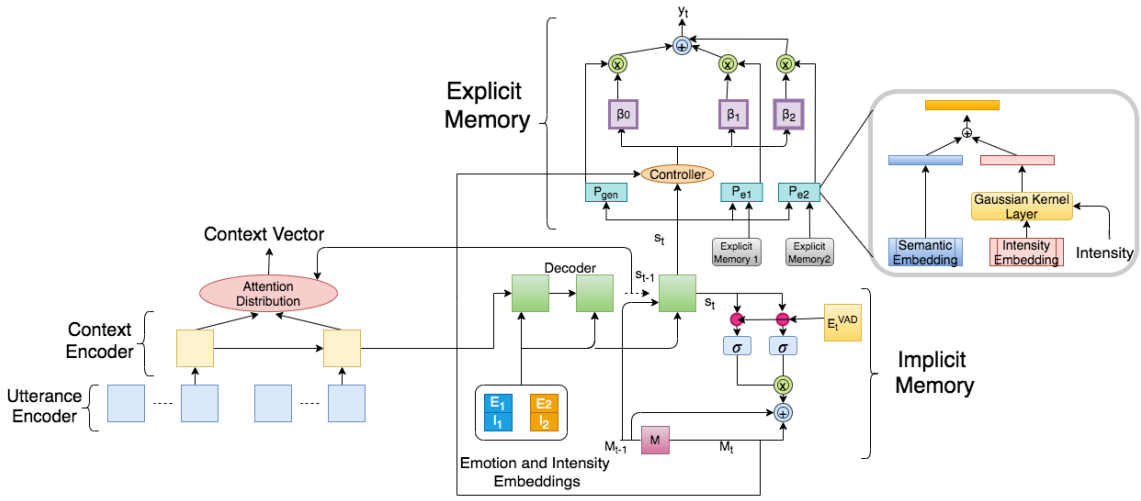

Figure 3: Architectural diagram of MEI-DG having a hierarchical encoder and decoder with explicit and implicit memory

V_{e1}, V_{e2} with the corresponding intensity I_1, I_2 and the previously decoded words. We use randomly initialized embedding to represent the desired emotion labels. Global Attention mechanism (Luong, Pham, and Manning 2015) is incorporated to enhance the performance of the decoder. The attention layer is applied to the hidden state of context encoder using decoder state d_t as the query vector. The concatenation of the context vector and the decoder state is used to compute the final probability distribution over the output tokens.

$$\begin{aligned} s_{d,t} &= LSTM_d(y_{t-1}, [s_{d,t-1}, c_t, V_{e_1}, V_{e_2}, \\ &\qquad I_1, I_2, M_{t-1}]) \\ c_t &= \sum_{i=1}^{k} \alpha_{t,i} s_{c,i}, \\ \alpha_{t,i} &= softmax(s_{c,i}^T W s_{d,t-1}) \end{aligned} \tag{3}$$

Implicit Memory: The emotion information in the form of embeddings provided to the decoder for generating emotional responses is static in nature. Thereby it fails to model the emotional dynamics. As discussed in (Ghosh et al. 2017), the above mentioned static approach decreases the fluency of the generated emotional response. Hence, in (Zhou et al. 2018), internal memory is designed to capture the emotional dynamics during decoding while maintaining fluency. Similarly, we incorporate an implicit memory to model the dynamics of the multiple emotions at the time of decoding. Before starting the decoding process, every conditioned emotion is assigned with an internal state initialized to 0, which at every step increases by a certain amount. Unlike (Zhou et al. 2018), the increment is measured by the amount of a given emotion left to be expressed in the generated response using VAD embeddings. At the end of decoding, the internal state should be equal to 1, indicating that the given emotion is completely expressed in the generated sentence. At every decoding step, to update the implicit memory state, we design an increase operation inc that decides the amount to be increased based upon the amount of emotion left to be expressed. A gate g_t is designed to control the usage of inc operation. Specifically, the internal state is updated as follows:

$$M_{e_i,t} = M_{e_i,t-1} + g_t(i) \circ \delta_t^{inc}(i) \tag{4}$$

where, $\circ$ is the element-wise multiplication and $g_t(i)$ and $\delta_t^{inc}(i)$ are the i^{th} element of g_t and δ_t^{inc}, respectively. $M_{e_i,t}$ is the implicit memory of the given i^{th} emotion at time-step t. Hence, the g_t and δ_t^{inc} operations are computed as:

$$\begin{aligned} g_t(i) &= \sigma(W_g[s_{d,t}; E_{i,t}^{VAD}] + b_g) \\ \delta_t^{inc}(i) &= \sigma(W_{inc}[s_{d,t}; E_{i,t}^{VAD}] + b_{inc}) \end{aligned} \tag{5}$$

where, $W_g, b_g, W_{inc}, b_{inc}$ are the trainable parameters; and $E_{i,t}^{VAD}$ represents the amount of emotion left at a time-step t for the emotion e_i.

$$E_{i,t}^{VAD} = E_{i,t-1}^{VAD} - y_{t-1}^{VAD} - \sum_{j=1, j\neq i}^{2} E_j \tag{6}$$

where, y_{t-1}^{VAD} is the Valence-Arousal-Dominance (VAD) embedding of the word generated at $t-1$ time-step. Here, E_j is the VAD embedding of a remaining set of emotions. The information of other emotions present in the generated sentence is subtracted to measure the amount of that particular emotion left to be expressed. For example, the given emotions are *surprise and joy*, and hence to incorporate the remaining amount of *surprise* in the response, we remove the information of *joy* from the given utterance to focus more on the emotion (*surprise*) left to be expressed in the response. To express surprise for a particular intensity the model should generate "Oh my God" instead it has only generated "Oh", then the implicit memory assists in incorporating the remaining amount of *surprise* emotion left to be expressed by generating "my God!". This facilitates in completely expressing a given emotion in the generated response.

Explicit Memory: Affect words, such as "great, wonderful", etc., provide strong emotional content to a response compared to generic words such as "book, night". Therefore, for expressing a given emotion, the model needs to have the ability to determine the trade-off between the generation of emotional words and generic words. Hence, two different sets of vocabularies are used to decide which words should be generated at a given time-step according to the specified emotion for generating the emotional response.

$$P(y_t) = \beta_0 P_{gen}(y_t) + \sum_{i=1}^{2} \beta_i P_{e_i}(y_t) \tag{7}$$

Here, P_{gen} is the probability of generating a generic word using the generic vocabulary computed as:

$$P_{gen}(y_t) = softmax(W_g s_{d,t}) \tag{8}$$

Though the similar idea of emotion versus generic words trade-off was considered previously (Zhou et al. 2018; Song et al. 2019), the authors did not account for intensity in emotions. For example, "wonderful, good, nice" all belong to the emotional category "joy", yet are different with respect to the intensity of the given emotion. Hence, we also incorporate the intensity factor and the choice between emotional or generic word at a given time-step of the decoder. To include the intensity-based emotional words, we calculate two sets of probabilities for a given target word at every decoder-step as:

$$P_{e_i}(y_t) = \gamma P_R(y_t) + \kappa P_I(y_t) \tag{9}$$

where, γ and κ are the learnable parameters and $P_R(y_t)$ signifies the relevance based generation probability with respect to the contextual history. Precisely, $P_R(y_t)$ can be described as follows:

$$P_R(y_t) = softmax(W_s s_{d,t}) \tag{10}$$

where W_s is the learnable parameter and $P_I(y_t)$ measures the generation probability of the target word given the target emotion intensity I_i. We assume that every word has intensity embedding beyond its semantic embedding, which stands for its preference under different intensities. The corresponding VAD embedding initializes the intensity embedding of every word. In VAD embedding, arousal is analogous to the intensity of a particular emotion. For example, the VAD embedding of "great" is [0.95 0.66 0.81], while of "good" is [0.94 0.37 0.54]. Hence, we can conclude that "great" has a higher intensity in comparison to "good". We employ a Gaussian Kernel Layer (Luong, Pham, and Manning 2015; Zhang et al. 2018; Luo et al. 2019) to encourage the words with intensity near the target intensity to be selected for generation. For example, if the specified target intensity for *sadness* is 0.9 then the words chosen should be "grief, misery, agony" than just "unhappy, sad" having lesser intensity.

$$\begin{aligned} P_I(y_t) &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\phi_I(UW) - I_i)^2}{2\sigma^2}) \\ \phi_I(U, W) &= \sigma(W^T(U \cdot W_{e_i} + b_{e_i})) \end{aligned} \tag{11}$$

where W is a one-hot indicator vector of the target word and ϕ_I maps the intensity embedding into real value, and the specified emotion intensity I_i is the mean of the Gaussian distribution. Here, σ^2 is the variance while W_{e_i} and b_{e_i} are the trainable parameters.

To ensure the incorporation of every conditioned emotion, we regulate β_0, β_1, and β_2 according to the implicit memory states that store the information of how much a particular emotion has been expressed in the generated response. If a certain amount of conditioned emotion is expressed, it tries to generate the remaining portion of the sequence such that the given emotion is completely expressed. For example, if "joy" is not completely expressed (i.e., the implicit memory state for *joy* doesn't change to 1), then the memory would select words that express this emotion. Also, if one of the emotions is expressed, the model needs to generate for the other emotions. For example, after "joy" getting completely expressed, the model focuses on generating other emotions like "surprise, sadness, etc". Precisely, β_0,β_1,β_2 are computed as:

$$\beta_0, \beta_1, \beta_2 = \sigma(W_p[s_{d,t}; M_{e_i,t}] + b_p) \tag{12}$$

where, W_p and b_p are the trainable parameters.

Training and Inference: The model is trained using the cross-entropy loss between the predicted response and the gold response, where the gold response is defined by $y^* = \{y_1^*, y_2^*, \ldots, y_m^*\}$. Apart from the cross-entropy loss, we add two regularization terms on the implicit and explicit memory. In implicit memory, the regularization term enforces the state to reach 1, indicating the desired emotion to be expressed entirely. Whereas, in explicit memory, it supervises the selection of the generic or word from a particular emotion. Finally, the loss is computed as:

$$\begin{aligned} \mathcal{L} = -\sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*) + \sum_{i=1}^{2} \|M_{e_i,n'}\| \\ - \sum_{k=0}^{2} o_k log(\beta_k) \end{aligned} \tag{13}$$

where $M_{e_i,n'}$ is the implicit emotion state at the last step M, o_k is true selection between the emotion and generic word and β_k is the probability of choosing a generic or emotion word.

Baseline Models: As mentioned before, this is one of the very first attempts that considers controlling multiple emotions along with their intensity for generating emotional responses in a dialogue setting. Hence, we did not find any multi-emotion with intensity baselines in the literature. ECM (Zhou et al. 2018), EMOTICONS (Colombo et al. 2019), and EmoDS (Song et al. 2019) are the suitable baselines for single-emotion. Therefore, we use these existing baselines to compare in case of single emotions only (for single emotion, we consider the higher intensity emotion for generation). For single emotion with intensity, we compare with Affect-LM(Ghosh et al. 2017) framework to witness the effectiveness of incorporating intensity values with emotion labels. For our multi-emotion baselines, we

implement the following models: **(i)**HRED: A general hierarchical encoder-decoder framework (Serban et al. 2015) and **(ii)** HRED+emb: Similar to (Song et al. 2019), we provide the multiple emotion categories in the form of vectors along with the corresponding intensity at each time-step of the decoder.

Experiments

Implementation details All the implementations are done using the Tensorflow (Abadi et. al.,2016) framework. For all the models including baselines, the batch size is set to 32. The utterance encoder is a bidirectional GRU with 600 hidden units in each direction. The context encoder and decoder are both GRUs with 600 hidden units. All the model parameters are randomly initialised using a Gaussian distribution with Xavier scheme(Glorot and Bengio, 2010). We employ AMS-Grad (Reddi et al., 2019) as the optimizer for model training to mitigate the slow convergence issues. The implicit memory is a trainable matrix of size 8300. To reduce data sparsity all the numbers and names are replaced with <number> and <person>. All out-of-vocabulary(OOV) words are replaced with <UNK> token. The generic vocabulary is a list of 40,000 words and emotion words (but emotion words have different markers). We use 300-dimensional word-embedding initialised with Glove (Pennington et al., 2014) embedding pre-trained on Twitter. We use 0.45 as dropout rate and perform gradient clipping when gradient norm is over 3. Previous 3 turns are considered for dialogue history and maximum utterance length is set to 50. The variance σ^2 of Gaussian Kernel Layer is set as 1. We ran 30 epoches, and proposed model took about a 1.5 week on a Titan X GPU machine.

Automatic Evaluation Metrics: We adopt perplexity, macro-average weighted F1 score and Pearson correlation co-efficient (Mohammad and Bravo-Marquez 2017) to evaluate the generated responses at content and emotional level, respectively, in a similar manner as (Akhtar et al. 2019). We also report embedding scores-based metrics (average, greedy, extreme) (Liu et al. 2016) similar to (Song et al. 2019).

(ii) Human Evaluation Metrics: We recruit six annotators from a third party company, having high-level language skills. We sample 250 responses per model for evaluation with the specified emotion and intensity provided for generation. First, we evaluate the quality of the response on two conventional criteria: *Fluency* and *Relevance*. These are rated on a five-scale, where 1, 3, 5 indicate unacceptable, moderate, and excellent performance, respectively, while 2 and 4 are used for unsure. Secondly, we evaluate the emotional quotient of a response in terms of both specified emotion and intensity. For the *Emotion* metric, the annotators are asked to judge whether the emotional category of the generated response is consistent with the specified emotions and the dialogue history. In the case of *Intensity* metric, they are asked to decide whether the degree of a particular emotion expressed in the generated response is according to the intensity specified for the given emotion. Here, 0 indicates irrelevant or contradictory, and 1 indicates consistent with the provided emotions or intensity values. We compute Fleiss' kappa (Fleiss 1971) to measure inter-rater consistency. The Fleiss' kappa for Fluency and Relevance are found to be 0.53 and 0.49, respectively, indicating moderate agreement. For Emotion and Intensity, we obtain 0.65 and 0.58 as the kappa scores indicating substantial agreement.

Models	PPL	Embedding			Emotion Content	
		Average	Greedy	Extreme	E-F1	IP-Corr
No Emotion						
HRED	80.7	0.491	0.360	0.371	0.39	0.26
Single Emotion						
HRED + emb	75.2	0.493	0.361	0.373	0.61	-
ECM (Zhou et al. 2018)	74.6	0.519	0.375	0.381	0.63	-
EMOTICONS (Colombo et al. 2019)	74.3	0.523	0.381	0.385	0.63	-
EmoDS (Song et al. 2019)	74.1	0.526	0.389	0.387	0.65	-
MEI-DG (Ours)	73.9	0.533	0.409	0.399	**0.67**	-
Single Emotion + Intensity						
HRED + emb	75.2	0.493	0.361	0.373	0.63	0.44
Affect-LM (Ghosh et al. 2017)	73.1	0.526	0.389	0.387	0.66	0.50
MEI-DG (Ours)	72.7	0.544	0.419	0.411	**0.69**	**0.57**
Multiple Emotion + Intensity						
HRED + emb	73.2	0.498	0.369	0.376	0.57	0.41
HRED + IM	72.9	0.512	0.396	0.413	0.59	0.48
HRED + EM - GK	74.1	0.531	0.412	0.407	0.60	0.43
HRED + EM	73.6	0.539	0.428	0.415	0.62	0.51
MEI-DG(HRED+EM+IM)	**71.2**	**0.552**	**0.443**	**0.428**	0.66	0.54

Table 3: Results for multiple emotions with intensity. We report PPL: Perplexity, Embedding scores, and Emotional content in terms of E-F1: macro-average weighted F1 score; IP: Intensity Prediction (Pearson correlation co-efficient). HRED+EM-GK denotes the model having explicit memory without the Gaussian Kernel(GK), and HRED+EM: denotes the model having explicit memory with Gaussian Kernel

Models	Fluency	Relevance	Emotion	Intensity
No Emotion				
HRED	3.17	2.89	15.9%	13.6%
Single Emotion				
HRED + emb	3.25	2.93	28.3%	-
ECM (Zhou et al. 2018)	3.45	3.08	36.7%	-
EMOTICONS (Colombo et al. 2019)	3.48	3.05	37.5%	-
EmoDS (Song et al. 2019)	3.47	3.12	39.2%	-
MEI-DG (Ours)	3.49	3.13	45.1%	-
Single Emotion + Intensity				
HRED + emb	3.52	3.21	32.5%	30.6%
Affect LM (Ghosh et al. 2017)	3.58	3.26	41.7%	34.2%
MEI-DG (Ours)	3.60	3.28	**47.3%**	**40.1%**
Multiple Emotion + Intensity				
HRED + emb	3.66	3.33	35.4%	32.9%
HRED + IM	3.71	3.40	40.7%	35.4%
HRED + EM - GK	3.69	3.38	38.6%	33.1%
HRED + EM	3.73	3.41	42.9%	36.8%
MEI-DG(HRED+EM+IM)	**3.82**	**3.48**	43.9%	37.2%

Table 4: Human evaluation results for all the baseline and proposed framework

Results and Discussion

We report the evaluation results in Table 3 and Table 4. We perform a comparative study on the MEIMD dataset, by providing single as well as multiple emotions with intensity. From the table, it can be seen that the models conditioned with single emotion outperform the model having no emotion information while the models conditioned with multiple emotions perform significantly better on all the metrics in case of both automatic and human evaluation. In Table 3 for response quality, our proposed MEI-DG framework conditioned with multiple emotions and their corresponding intensity values have a lower perplexity of 71.2 than all the other baseline methods. Additionally, our embedding metrics are significantly better than all the existing baselines de-

Conversational History	Model	Emotion	Intensity	Generated Response
It was an amazing night. You think it looked amazing?	HRED	-	-	Yes, I think so.
	MEI-DG	Joy	0.6	Yes it was really fascinating, I enjoyed it.
	MEI-DG	Joy	0.3	Yes it was a nice night.
	MEI-DG	Sadness	0.3	I found the night gloomy.
	MEI-DG	Sadness	0.9	The whole night I cried all alone.
	MEI-DG	Joy, Sadness	0.3, 0.6	It was nice but I terribly missed my friend.
	MEI-DG	Joy, Sadness	0.9, 0.3	The night was really great only had a little trouble in finding cab.
I invented the game of Cups as a way to give Joey money. And now you want that money back.	HRED	-	-	I want it back.
	MEI-DG	Anger	0.6	What the hell is wrong with you?
	MEI-DG	Anger	0.9	I would kill you if I don't get my money back.
	MEI-DG	Surprise	0.3	Wow, I didn't know about it.
	MEI-DG	Surprise	0.9	Oh my God! I left the money in the room.
	MEI-DG	Anger, Surprise	0.3, 0.6	It's not a game! Stop it and give me the money.
	MEI-DG	Anger, Surprise	0.6, 0.3	Oh you please stop annoying me.

Table 5: Examples of responses generated by MEI-DG that are conditioned on single and multiple emotions with intensity.

signed to evaluate the relevance of the generated response.

The F1 score for emotion classification and Pearson correlation coefficient for intensity prediction for the proposed MEI-DG framework with responses conditioned on multiple emotions is less than the responses being conditioned with a single emotion. This performance reduction can be attributed to the fact that unlike dealing with single-emotion, handling multiple emotions is comparatively a harder task. Although there is a decrease in performance, in our multiple emotion-based MEI-DG framework, we see an absolute improvement of 3% in emotion F1-score in contrast to the existing ECM, EMOTICONS framework. Also, there is an increase of 4% in intensity prediction score compared to the Affect-LM network. This validates the fact that the proposed multi-emotion framework is capable of generating overall better responses. We perform ablation tests to quantify the contributions made by the different components, as shown in Table 3. From the ablation study results, we can conclude that the intensity performance of MEI-DG decreases in the absence of a Gaussian kernel that controls the intensity factor in the generated response. With the addition of Gaussian kernel along with the explicit memory *HRED+EM*, there is a significant improvement of 8% in the performance of our proposed framework compared to the *HRED+EM-GK* network. The final model, having both implicit and explicit memory, gives the best results than all the variants (*HRED+IM, HRED+EM-GK, HRED+EM*) of the proposed model.

The results of the human evaluation are presented in Table 4. Compared to the existing frameworks, our proposed model obtains the highest emotion and intensity score of 43.9% and 37.2%, respectively. The single emotion and intensity-based MEI-DG model has the highest score in terms of emotion and intensity instead of all the baseline and existing networks, and outperforms the final proposed framework, aligning with the automatic evaluation results. Though the emotion and intensity scores are less in multiple emotion models than single emotion models, the fluency and relevance scores are better than the single emotion frameworks. One plausible explanation is that the responses are more coherent and relevant in expressing the speaker's complete emotional state.

We provide some generated examples from the multiple emotion MEI-DG model and single emotion and no emotion baselines in Table 5. From the table, it is evident that our proposed framework is capable of inducing multiple emotions in the generated response (*joy and sadness* or *anger and surprise*). Besides, the generated responses can variate the different levels of intensity (*0.3, 0.6, 0.9*) for a particular emotion. For different emotions with the variation in intensity the choice of words changes, for example, low intensity *surprise* is expressed by "wow" while high intensity is expressed by the phrase "Oh my God!". This validates the fact that the proposed network has fair understanding of the different levels of intensity for generation. Finally, in comparison to the single emotion models, the multi-emotion framework generates more comprehensive and emotionally complete responses relevant to the given dialogue context. After performing qualitative analysis, we came across some of the commonly occurring errors committed by MEI-DG that is at times MEI-DG is incapable of generating responses having emotions such as "joy, anger", "surprise, sadness", "acceptance, disgust" due to fewer occurrence of these emotions simultaneously in an utterance.

Conclusion and Future Work

In this work, we give a methodology of generating responses conditioned on multiple emotions and their corresponding intensity. We have created a large-scale dataset- MEIMD-labeled with multiple emotions and the corresponding intensity.We have designed a framework- MEI-DG-, comprising of an implicit memory that regulates the amount of emotions left to be expressed in the generated response. The explicit memory finds a trade-off between the words to be generated at a given time-step. Also, the Gaussian kernel that is part of the explicit memory, promotes generation of emotional words having appropriate intensity. From the experimental results- both quantitative and qualitative- we can conclude that MEI-DG can generate emotional responses in accordance with the specified emotions and the desired intensity. As our approach is model agnostic, in future we would like to investigate the performance gain of the pre-trained transformers based architecture, like BERT and GPT-2 over RNN for dialogue modeling. We would also like to incorporate multimodal information along with text like audio and video to see the effectiveness over text-only.

Acknowledgements

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

Akhtar, S.; Ghosal, D.; Ekbal, A.; Bhattacharyya, P.; and Kurohashi, S. 2019. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing* .

Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective Neural Response Generation. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, 154–166. Springer.

Chauhan, D. S.; Akhtar, M. S.; Ekbal, A.; and Bhattacharyya, P. 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)*, 5651–5661.

Colombo, P.; Witon, W.; Modi, A.; Kennedy, J.; and Kapadia, M. 2019. Affect-Driven Dialog Generation. *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)* 3734–3743.

Ekman, P.; Friesen, W. V.; O'sullivan, M.; Chan, A.; Diacoyanni-Tarlatzis, I.; Heider, K.; Krause, R.; LeCompte, W. A.; Pitcairn, T.; Ricci-Bitti, P. E.; et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53(4): 712.

Firdaus, M.; Chauhan, H.; Ekbal, A.; and Bhattacharyya, P. 2020a. EmoSen: Generating Sentiment and Emotion Controlled Responses in a Multimodal Dialogue System. *IEEE Transactions on Affective Computing* .

Firdaus, M.; Chauhan, H.; Ekbal, A.; and Bhattacharyya, P. 2020b. MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4441–4453.

Firdaus, M.; Thangavelu, N.; Ekba, A.; and Bhattacharyya, P. 2020c. Persona aware Response Generation with Emotions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

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

Ghosh, S.; Chollet, M.; Laksana, E.; Morency, L.-P.; and Scherer, S. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers* 634–642.

Huang, C.; Trabelsi, A.; and Zaïane, O. R. 2020. Seq2Emo for Multi-label Emotion Classification Based on Latent Variable Chains Transformation. *AAAI* .

Huang, C.; Zaiane, O. R.; Trabelsi, A.; and Dziri, N. 2018. Automatic Dialogue Generation with Expressed Emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 49–54.

Kim, Y.; Lee, H.; and Jung, K. 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* 141–145.

Li, J.; and Sun, X. 2018. A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018* 678–683.

Li, Q.; Chen, H.; Ren, Z.; Chen, Z.; Tu, Z.; and Ma, J. 2020. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation. *AAAI* .

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers* 986–995.

Lin, Z.; Xu, P.; Winata, G. I.; Liu, Z.; and Fung, P. 2019. CAiRE: An End-to-End Empathetic Chatbot. *arXiv preprint arXiv:1907.12108* .

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016* 2122–2132.

Lubis, N.; Sakti, S.; Yoshino, K.; and Nakamura, S. 2018. Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5293–5300.

Luo, F.; Dai, D.; Yang, P.; Liu, T.; Chang, B.; Sui, Z.; and Sun, X. 2019. Learning to Control the Fine-grained Sentiment for Story Ending Generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 6020–6026.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* 1412–1421.

Martinovski, B.; and Traum, D. 2003. Breakdown in human-machine interaction: The error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, 11–16.

McHugh, M. L. 2012. Interrater reliability: The Kappa statistic. *Biochemia medica: Biochemia medica* 22(3): 276–282.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2322–2332. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1255. URL https://www.aclweb.org/anthology/D18-1255.

Mohammad, S. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20, 000 English Words. 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*, 174–184.

Mohammad, S. M.; and Bravo-Marquez, F. 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* 34–49.

Plutchik, R. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89(4): 344–350.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 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* 527–536.

Prendinger, H.; Mori, J.; and Ishizuka, M. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies* 62(2): 231–245.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 5370–5381.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2015. Hierarchical Neural Network Generative Models for Movie Dialogues. *arXiv preprint arXiv:1507.04808* 7(8).

Shen, L.; and Feng, Y. 2020. CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation. *arXiv preprint arXiv:2005.00329* .

Song, Z.; Zheng, X.; Liu, L.; Xu, M.; and Huang, X.-J. 2019. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 3685–3695.

Winata, G. I.; Madotto, A.; Lin, Z.; Shin, J.; Xu, Y.; Xu, P.; and Fung, P. 2019. CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification. *arXiv preprint arXiv:1906.04041* .

Yeh, S.-L.; Lin, Y.-S.; and Lee, C.-C. 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*, 6685–6689. IEEE.

Yu, J.; Marujo, L.; Jiang, J.; Karuturi, P.; and Brendel, W. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. 1097–1102.

Zhang, R.; Guo, J.; Fan, Y.; Lan, Y.; Xu, J.; and Cheng, X. 2018. Learning to Control the Specificity in Neural Response Generation. 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*, 1108–1117.

Zhong, P.; Wang, D.; and Miao, C. 2019. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 7492–7500.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 730–739.