SEPRG: Sentiment aware Emotion controlled Personalized Response Generation

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

Abstract

Social chatbots have gained immense popularity, and their appeal lies not just in their capacity to respond to the diverse requests from users, but also in the ability to develop an emotional connection with users. To further develop and promote social chatbots, we need to concentrate on increasing user interaction and take into account both the intellectual and emotional quotient in the conversational agents. Therefore, in this work, we propose the task of sentiment aware emotion controlled personalized dialogue generation giving the machine the capability to respond emotionally and in accordance with the persona of the user. As sentiment and emotions are highly co-related, we use the sentiment knowledge of the previous utterance to generate the correct emotional response in accordance with the user persona. We design a Transformer based Dialogue Generation framework, that generates responses that are sensitive to the emotion of the user and corresponds to the persona and sentiment as well. Moreover, the persona information is encoded by a different Transformer encoder, along with the dialogue history, is fed to the decoder for generating responses. We annotate the PersonaChat dataset with sentiment information to improve the response quality. Experimental results on the PersonaChat dataset show that the proposed framework significantly outperforms the existing baselines, thereby generating personalized emotional responses in accordance with the sentiment that provides better emotional connection and user satisfaction as desired in a social chatbot.

1 Introduction

One of the significant challenges of artificial intelligence (AI) is to endow the machine with the ability to interact in natural language with humans. In the recent past, tremendous effort has been given to create smart personal assistants, such as Microsoft's Cortana, Apple's Siri, Amazon's Alexa, Google Home, etc. The personal assistants in our mobile devices invariably assist in our day-to-day lives by answering a wide range of queries. Such assistants act as social agents that take care of the various activities of their users. Besides reacting passively to user requests, they also proactively anticipate user needs and provide in-time assistance, such as reminding of an upcoming event or suggesting a useful service without receiving explicit user requests (Sarikaya, 2017). The daunting task for these agents is that they have to work well in open domain scenarios as people learn to rely on them to effectively maintain their works and lives efficiently.

Empathy and social belonging are a few of the fundamental needs for human beings (Maslow, 1943). Building social chatbots to tackle these emotional needs is indeed of great benefit to our society. The primary objective of these chatbots is not inherently to answer all the users' questions, but rather to be a virtual companion of the users. To become a better companion, it is imperative for the agent to understand the personality of the user to assist in different aspects of life. Along with understanding the personality traits of a user, the emotional connection is an essential factor for building better communication. Social conversational agents can serve for a more extended period of time by maintaining a consistent personality (increasing trust in the user) and by establishing an emotional connection with them. The ability to empathize, create social belonging, and adhere to a personality and integration of these factors in conversational agents is one of the long-standing goals of Artificial Intelligence (AI). Conversational agents need to monitor the user's emotion in order to suffice the emotional needs and simultaneously empathize with them, making the conversation engaging, increasing user contentment (Prendinger et al., 2005),

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 353–363,
Aberdeen, Scotland, UK, 20-24 September 2021.

Persona 1	Persona 2
I am primarily a meat eater.	*I've a sweet tooth.*
I am a guitar player.	*I'm a babysitter and drive a mercedes.*
Welding is my career field.	*I'm the middle child of 3 siblings.*
My parents don't know I am gay.	*I'm getting married in six months.*
[Person 1] What do you do for career? (Neutral)	
[Person 2] I like to watch kids. (Positive)	
[Person 1] I actually play guitar and do a lot of welding. (Positive)	
[Person 2] What do you weld? houses?(Neutral)	

Table 1: A conversation from the PersonaChat dataset with sentiments

and decreasing breakdowns in conversations (Martinovski and Traum, 2003). Moreover, these agents should also have the capability to generate personalized responses conforming to the personal interests and unique needs of different users while presenting a consistent personality to gain the user's trust and confidence. Hence, the primary motivation of our current work lies in generating responses that are engaging, emotionally appropriate, and also integrates the personal interests of the user.

Lately, researchers have started focusing on incorporating personality information on chit-chat (Zhang et al., 2018) and goal-oriented (Joshi et al., 2017; Luo et al., 2019) conversational systems. Due to the lack of persona data sets, the authors created a PersonaChat dataset in (Zhang et al., 2018), where the individual personality data is represented in a few texts for open-domain chit-chat dialogue systems. We present an example from the dataset in Table 1, from which it is obvious that the speakers are able to retain the persona knowledge when communicating with each other. This helps to make the dialogue engaging and also makes it easier to build trust and credibility with the users (Shum et al., 2018). For conversational systems to effectively communicate with the user in a coherent and natural way, the ability to maintain a clear persona is imperative. While it is necessary to maintain a clear personality in order to gain the confidence of the user, it is also essential to react emotionally in order to create a bond with the user.

From Table 1, it is evident that when talking with the user, the agent can retain a specific personality, but it sacrifices the emotional link with the user. The dialogue, therefore, is almost like stating facts instead of a real discussion. In this work, therefore, we propose the task of infusing the responses with emotional content while maintaining a clear persona. From the table, the response to *Person 2* could be more empathetic like *That's a great job, as I play guitar and do welding for a career.* This response has a happy undertone than the ground-truth response, which is neutral and contains only the facts about *Person 1*. Empathetic responses are insightful and provide a forum for a more substantial discussion. It is evident from the illustration that only having a persona in a reply is not sufficient to produce interactive responses. To render it more human-like, the emotional element must also be integrated into the replies. 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. Every emotion is associated with sentiments, hence using the sentiment information of the utterances can assist in narrowing down the set of emotions for generating contextually correct emotional responses. In the Table 1, the dialogue has been annotated with the corresponding sentiments to assist in generating empathetic responses. To the best of our knowledge, this is one of the first works that include sentiment information for creating personalized emotional responses.

The key contributions of this work are as follows:

1. We propose the task of generating empathetic, personalized responses while considering the persona information and implicitly the sentiment in the responses through the dialogue context.

2. We propose a novel Transformer based encoder-decoder framework, with the ability to infuse the sentiment, emotion and persona information in the responses.

3. Experimental results show that our proposed framework is capable of maintaining a consistent persona and sentiment while generating emotional responses compared to the existing baselines.

The rest of the paper is structured as follows. In Section II, we present a brief survey of the related work. In Section III, we explain the proposed

methodology. In Section IV, we describe the details of the datasets that we used and annotated. The experimental setup, along with the evaluation metrics, is reported in Section V. In Section VI, we present the results along with the necessary analysis. Finally, we conclude in Section VII with future work.

2 Related Work

In complete applications, such as dialogue systems, natural language generation (NLG) has become increasingly essential (Vinyals and Le, 2015; Li et al., 2016b; Serban et al., 2017; Wu et al., 2018) and also in many other natural language interfaces. The generation of responses provides the means by which a conversational agent can communicate with its user to assist users in achieving their desired goals. Recently, generative adversarial networks have been exploited for dialogue generation (Xu et al., 2018, 2017; Zhang et al., 2019; Zhu et al., 2019; Bruni and Fernandez, 2017) for a better generation of responses.

Persona information is an essential part of generating responses. Earlier works on persona-based conversational models (Li et al., 2016a) incorporated speakers' embeddings to infuse persona information in the responses. To incorporate persona in chit-chat models, the authors in (Zhang et al., 2018; Mazaré et al., 2018) introduced a PersonaChat dataset that includes personal information of the speakers. This dataset has been extensively used to build persona-based dialogue systems (Madotto et al., 2019; Yavuz et al., 2019; Song et al., 2019, 2020). The authors in (Madotto et al., 2019) used a meta-learning framework to include persona information in the generated responses. Similarly, the authors in (Yavuz et al., 2019) employed a hierarchical pointer network for generating persona-based responses. The authors in (Song et al., 2019) used persona information to generate diverse responses by employing conditional variational autoencoder. Our present work differs from these existing works (that made use of the PersonaChat dataset) in a sense that we intend to use the persona information while generating emotional responses.

Persona information is also being exploited in goal-oriented dialogue systems (Joshi et al., 2017; Luo et al., 2019; Qian et al., 2017). The authors in (Joshi et al., 2017) introduced persona information in the babI dialog dataset for creating better responses. The authors used conditional variational auto-encoders for personalized generation in (Wu et al., 2020). As personalization has been considered in responses, we intend to take a step ahead by inculcating the emotions in accordance to the emotion of the user and the dialogue history.

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; Kezar, 2018; Rashkin et al., 2019; Zhou and Wang, 2017). In (Zhou et al., 2018), an emotional chatting machine (ECM) was proposed that was built upon seq2seq framework for generating emotional responses. Recently, a lexicon-based attention framework was employed to generate responses with a specific emotion (Song et al., 2020). Emotional embedding, along with affective sampling and regularizer, was employed to generate the affect driven dialogues in (Colombo et al., 2019). Lately, authors in (Firdaus et al., 2020) designed personalized response generation framework with controllable emotions using basic sequence-to-sequence framework. Our present research differs from these existing works as we propose a novel framework using a generative adversarial network to generate responses in an empathetic manner, having a consistent persona.

3 Methodology

We define the problem statement in this section, followed by the detailed descriptions of the proposed methodology. The architectural diagram of the sentiment and persona guided emotional dialogue generation framework is presented in Figure 1.

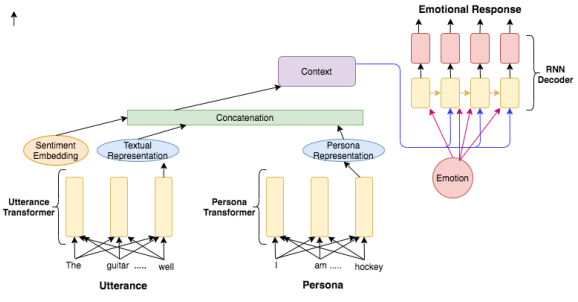

Figure 1: Architectural diagram of the proposed framework.

Problem Definition: In our present work we aim at solving the task of emotional and personalized dialogue generation in accordance to the conversational history, sentiment and the persona information of the speaker. For a given sequence

of dialogue turns $D = [U_1, U_2, \ldots, U_N]$ as the dialogue context, where $U_n = [w_1, w_2, \ldots, w_k]$ is the n^{th} dialogue turn and each dialogue turn is associated with sentiment labels represented by $S_{lab} = s_1, s_2, \ldots, s_N$ having a set of persona information $P = P_1, P_2, \ldots, P_m$ the task is to generate an emotional personalized response $Y = y_1, y_2, \ldots, y_{n'}$ along with the emotion embedding V_e for the desired emotion E that is sensitive to the speaker's expressed sentiment and is consistent to the persona information.

3.1 Proposed Framework

Our proposed framework is based upon the Transformer network (Vaswani et al., 2017) as shown in Figure 1. Our network comprises of two encoders: an utterance encoder to transform the textual utterance $U_i = (w_{k,1}, w_{k,2}, ..., w_{k,n})$ and a persona encoder to encode the set of persona information $P = P_1, P_2, \ldots, P_m$. Finally, we employ a transformer decoder to generate emotionally controlled responses according to the specified emotions in a similar manner as (Firdaus et al., 2020; Huang et al., 2018; Zhou et al., 2018).

Utterance Encoder: As the transformer encoder has multiple layers and each layer is composed of a multi-head self attentive sub-layer followed by a feed-forward sub-layer with residual connections (He et al., 2016) and layer normalization (Ba et al., 2016), we use it to encode the utterances in a given dialog. For intricate details on the Transformer network, we refer the interested readers to (Vaswani et al., 2017). To learn the representation of $U_i = (w_{k,1}, w_{k,2}, ..., w_{k,n})$ is first mapped into continuous space

$$T_u = (t_1^i, t_2^i, \ldots, t_{|U_i|}^i); where[T_j^i = e(w_j^i) + p_j] \quad (1)$$

where $e(u_j^i)$ and p_j are the word and positional embedding of every word u_j^i in an utterance, respectively. For words we use Glove embeddings and we adopt sine-cosine positional embedding (Vaswani et al., 2017) as it performs better and does not introduce additional trainable parameters. The utterance encoder (a Transformer) converts T_u into a list of hidden representations $h_1^i, h_2^i, \ldots, h_{|U_i|}^i$. We use the last hidden representation $h_{|U_i|}^i$ (i.e. the representation at the EOS token) as the textual representation of the utterance U_i. Similarly, to the representation of each word in U_i, we also take into account the utterance position. Therefore, the final textual representation of the utterance U_i is:

$$h_i^s = h_{|U_i|}^i + p_i \quad (2)$$

Note that the words and sentences share the same positional embedding matrix. We also concatenate the sentiment information of every sentences represented by $S_{lab} = s_1, s_2, \ldots, s_N$. The final representation of any utterance is given by the concatenation of the sentiment representation as well as the last hidden representation of the utterance.

$$h_i^{utt} = h_i^s + s_N \quad (3)$$

Persona Encoder: To learn the representation of the set of persona information $P = P_1, P_2, \ldots, P_m$ is first mapped into continuous space

$$T_p = (t_1^i, t_2^i, \ldots, t_{|P_m|}^i); where[T_k^i = e(w_k^i) + p'_k] \quad (4)$$

where $e(u_k^i)$ and p'_k are the word and positional embedding of every word u_k^i in a given persona, respectively. Similar to the utterance encoder, for words we use the Glove embeddings and adopt sine-cosine positional embedding (Vaswani et al., 2017). The persona encoder (a Transformer) converts T_p into a list of hidden representations ${h'_1}^i, {h'_2}^i, \ldots, {h'_{|P_m|}}^i$. We use the last hidden representation $h_{|P_m|}^i$ (i.e. the representation at the EOS token) as the persona representation of the given speaker. Therefore, the final persona representation of the utterance P_m is:

$$h_i^p = h_{|P_m|}^i + p'_i \quad (5)$$

Emotion controlled Decoder: To generate the next textual response with the given emotion information we employ a RNN decoder as shown in Figure 1. We employ GRU for generating the response in a sequential manner based on the context hidden representation from both the transformers, and the words decoded previously. We use the input feeding decoding along with the attention (Luong et al., 2015) mechanism for enhancing the performance of the model. Using the decoder state $h_{d,t}^{dec}$ as the query vector, we apply self-attention on the hidden representation of the utterance-level encoder. The decoder state, persona information and the context vector are concatenated and used to calculate a final distribution of the probability

over the output tokens.

$$
\begin{aligned}
h^{dec}_{d,t} &= GRU_d(y_{k,t-1}, h_{d,t-1}) \\
c_t &= \sum_{i=1}^{k} \alpha_{t,i} \hat{D}, \\
\alpha_{t,i} &= softmax(\hat{D}^T W_f h_{d,t}) \\
\tilde{h}_t &= tanh(W_{\tilde{h}}[h_{d,t}; c_t]) \\
P(y_t / y_{<t}) &= softmax(W_V \tilde{h}_t)
\end{aligned}
\tag{6}
$$

where, W_f, W_V and $W_{\tilde{h}}$ are the trainable weight matrices.

For generating responses with the specified emotion as shown in Figure 1, we provide the emotion vector V_e (the emotion embeddings are pre-trained Glove embeddings) as input during decoding at every decoder time-step. In order to include the emotion vector in the decoder, we modify Equation (6) to incorporate the emotion information for the generation of responses and the modified equation is as follows:

$$
h^{dec}_{d,t} = GRU_d(y_{k,t-1}, [h_{d,t-1}, V_e]) \tag{7}
$$

Training and Inference: We employ commonly used teacher forcing (Williams and Zipser, 1989) algorithm at every decoding step to minimize the negative log-likelihood on the model distribution. We define $y^* = \{y^*_1, y^*_2, \ldots, y^*_m\}$ as the ground-truth output sequence for a given input by:

$$
\mathcal{L}_{ml} = -\sum_{t=1}^{m} \log p(y^*_t | y^*_1, \ldots, y^*_{t-1}) \tag{8}
$$

Baseline Models: We develop the following baselines: (i) Seq2Seq: This is a basic encoder-decoder (Sutskever et al., 2014) framework with no persona, sentiment and emotion information. (ii). HRED: A general hierarchical encoder-decoder framework (Serban et al., 2017) that captures the conversational context without the persona, sentiment and emotion information. (iii). Seq2Seq + E + P: The utterance encoder along with persona encoder and emotion information is used to decode the responses in a similar manner as (Firdaus et al., 2020). (iv). HRED + E + P: We infuse the persona and emotion in the basic hierarchical encoder-decoder framework. (v). Seq2Seq + E + P + S: The utterance encoder along with persona encoder, sentiment information and emotion information is used to decode the responses. (vi). HRED + E + P + S: We infuse the persona, sentiment and emotion in the basic hierarchical encoder-decoder framework. (vii) Trans: Basic transformer network without persona, sentiment and emotion information. (viii) Trans + E + P: The transformer encoders along with persona encoder and emotion information is used to generate the responses.

4 Dataset and Experimentation

Dataset Description: On the recently published ConvAI2 benchmark dataset, which is an extended version (with a new test set) of the persona-chat dataset (Zhang et al., 2018), we conduct our experiments. The interactions are collected from the randomly paired crowd workers who were instructed to play the part of a given persona. In over 10,981 dialogues, this dataset comprises of 164,356 utterances and has a collection of 1,155 personas, each consisting of at least four personality texts. There are 1,016 dialogues in the testing set and 200 never before seen personas. As the dataset is not labeled with emotions, we use the emotion annotated version of the dataset used in (Firdaus et al., 2020).

Dataset Preparation: As sentiment and emotions are highly co-related we annotate the PersonaChat dataset using the emotion information in a similar manner as (Poria et al., 2019). As emotions such as *excited, grateful, joyful, caring, hopeful, faithful, impressed* have a positive undertone hence we automatically label the utterances having these emotion labels as positive sentiment. Similarly for emotions such as *angry, sad, annoyed, disgusted, terrified, furious, disappointed, jealous* has a negative undertone hence are labelled as negative sentiment. For the other emotion labels such as *surprise, proud, nostalgic, guilty, confident, prepared, sentimental* that can either be positive, neutral or negative depending on the utterance and the context we resort to manual annotation. For annotating the utterances in the PersonaChat 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. Majority voting scheme was used for selecting the final sentiment label for each utterance. We achieve an overall Fleiss' (Fleiss, 1971) kappa score of 0.75 for sentiment which can be considered as reliable. Detailed statistics of the PersonaChat dataset are provided in Table 2.

Implementation Details: All the implementa-

Dataset Statistics	Train	Valid	Test
Dialogues	7686	1640	1655
Utterances	124816	19680	19860
Avg. turns per Dialogue	12.51	12.73	12.74
Avg. words in a Response	11.89	9.57	10.75
Emotions per dialogue	7.4	6.5	5.1
Unique words	20322	13415	15781

Table 2: Statistics of the PersonaChat Dataset

tions are done using the PyTorch[1] framework. For all the models, including baselines, the batch size is set to 32. We use the dropout (Srivastava et al., 2014) with probability 0.45. During decoding, we use a beam search with beam size 10. We initialize the model parameters randomly using a Gaussian distribution with the Xavier scheme (Glorot and Bengio, 2010). We employ AMSGrad (Reddi et al., 2019) as the optimizer for model training to mitigate the slow convergence issues. We use uniform label smoothing with $\epsilon = 0.1$ and perform gradient clipping when the gradient norm is over 5. To reduce data sparsity, all the numbers and names are replaced with <number> and <person>.

Automatic Evaluation Metrics: In order to assess the model at the emotional and grammatical level, we present the results using the traditional automatic metrics. Perplexity(Chen et al., 1998) is stated to test our proposed framework at the content level. We also report the results using the standard metrics like BLEU-4 (Papineni et al., 2002) and Rouge-L (Lin, 2004) to measure the ability of the generated response for capturing the correct information. BLEU measures the n-grams overlap between the generated response and the gold response, and has become a standard measure for comparing task-oriented dialog systems. It is used to measure the content preservation in the generated responses. We report Distinct-1 and Distinct-2 metrics that measure the distinct n-grams in the generated responses and are scaled with respect to the total number of generated tokens to avoid repetitive and boring responses (Li et al., 2016b). To measure the emotional content in the generated responses, we calculate the emotion accuracy using the pre-trained BERT classifier on the responses generated by the baseline and proposed models.

Human Evaluation Metrics: We randomly sample 500 responses from the test set for human evaluation. For a given input along with persona information, six annotators with post-graduate exposure were assigned to evaluate the quality of the generated responses by the different approaches in a similar manner as the existing works (Firdaus et al., 2020). 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 persona, sentiment and emotion inclusion in response in terms of *Persona Consistency* metric, *Sentiment Coherence* metric and *Emotion Appropriateness* to judge whether the response generated is in consonance to the specified persona, sentiment and the emotion is also coherent to the conversational history. In the case of all these metrics, 0 indicates an irrelevant or contradictory persona, sentiment or emotion in the response, and 1 represents the consistent response to the specified persona, sentiment and emotion. For the human evaluation metrics, we calculate the Fleiss' kappa (Fleiss, 1971) to determine the inter-rater consistency. For fluency and relevance, the kappa score is 0.75, and for emotion appropriateness, sentiment coherence and persona consistency, these are 0.75, 0.71 and 0.78, respectively, indicating "substantial agreement".

5 Result and Analysis

For thorough analysis of our proposed framework, we provide a detailed analysis of the results (both automatic and manual) along with the generated responses. We also analyze the errors made by the network in generating empathetic personalized responses.

Automatic Evaluation Results: The automatic evaluation results are presented in Table 3, which demonstrates that the proposed framework significantly outperforms all the baselines with respect to all the metrics. The final proposed transformer network shows a notable drop in perplexity scores, thereby ensuring grammatically correct responses generated by the framework. In addition, we see that the BLEU scores have increased with an improvement of more than 5% from the basic Seq2Seq framework and with a gain of 4% from the typical HRED model. By introducing the persona and emotion information in the basic Seq2Seq and HRED model, we see the growth in performance, establishing the need for persona and emotion knowledge for generating empathetic, personalized responses. Similarly, in the case of Rouge-L,

[1]https://pytorch.org/

Model Description		Perplexity	BLEU-4	Rouge-L	Distinct-1	Distinct-2	Emotion Accuracy
Baseline Approaches	***Seq2Seq*** *(Sutskever et al., 2014)*	56.11	0.089	0.196	0.0125	0.0464	0.358
	HRED *(Serban et al., 2017)*	55.63	0.096	0.201	0.0128	0.0469	0.376
	Seq2Seq + E + P *(Firdaus et al., 2020)*	54.13	0.103	0.189	0.0168	0.0549	0.657
	HRED + E + P	54.85	0.116	0.224	0.0174	0.0592	0.665
	Seq2Seq + E + P + S	53.61	0.115	0.203	0.0171	0.0555	0.673
	HRED + E + P + S	52.46	0.127	0.237	0.0186	0.0590	0.689
Proposed Approach	***Trans + E + P + S***	**51.92**	**0.143**	**0.266**	**0.0219**	**0.0987**	**0.715**
Ablation Study	***Trans***	53.47	0.118	0.239	0.0189	0.0883	0.678
	Trans + E + P	53.44	0.125	0.242	0.0193	0.0896	0.695

Table 3: Results of automatic evaluation. Here, E-Emotion, P-Persona, S-Sentiment, Trans-Transformers

Model Description		Fluency	Relevance	Emotion Appropriateness	Persona Consistency	Sentiment Coherence
Baseline Approaches	***Seq2Seq*** *(Sutskever et al., 2014)*	2.98	2.65	38%	35%	33%
	HRED *(Serban et al., 2017)*	3.16	2.89	41%	39%	37%
	Seq2Seq + E + P *(Firdaus et al., 2020)*	3.29	3.02	53%	48%	42%
	HRED + E + P	3.45	3.18	59%	55%	47%
	Seq2Seq + E + P + S	3.33	3.21	56%	52%	51%
	HRED + E + P + S	3.52	3.45	61%	58%	55%
Proposed Approach	***Trans + E + P + S***	**3.89**	**3.66**	**65%**	**67%**	**62%**

Table 4: Results of human evaluation for the existing baselines and the proposed framework

Dialog Input	Persona Information		Ground-Truth	Generated Response
[Person 1] What do you do for a career? [Person 2] I like to watch kids. [Person 1] I actually play guitar and do lot of welding.	Persona 1 I am primarily a meat eater. I am a guitar player. Welding is my career field. My parents don't know I am gay.	Persona 2 I've a sweet tooth. I'm a babysitter and drive a mercedes. I'm the middle child of 3 sibling. I'm getting married in six weeks.	What do you weld houses? Neutral	**Seq2Seq:** What is weld? ¡UNK¿ ¡UNK¿ **HRED:** What you weld? **Seq2Seq+E+P:** Wow! What you weld mostly? (surprise) **HRED+E+P:** Really! Do you weld houses? (surprise) **Proposed:** Wow that's great, what do you weld mostly, is it houses? (surprise)
[Person 1] Hi! do you like turtles? [Person 2] I have two cats actually. [Person 1] I have a turtle his name is Speedy.	Persona 1 I don't pick up my toys. I have a pet turtle. I like to play with my dolls. I'm fourteen.	Persona 2 I love cats and have two cats. My favorite season is winter. I won a gold medal in the 1992 olympics. I've a hat collection of over 1000 hats.	I am a cat person. Positive	**Seq2Seq:** I have cats. **HRED:** I like cats mostly. **Seq2Seq+E+P:** Turtles are nice but I like cats. (joy) **HRED+E+P:** Nice name for a pet, but I love cats. (joy) **Proposed:** That is an adorable name for a turtle! but I am a cat person. (joy)

Table 5: Examples of responses generated by different models having emotion, sentiment and persona

we see a remarkable improvement in the performance of the proposed network compared to the *Seq2Seq+E+P* and *HRED+E+P* frameworks, respectively. We also report the emotion accuracy of the generated response. It is quite apparent that the responses having emotion information have higher accuracy than the models with no emotion information. The proposed model outperforms the best performing baseline network with an improvement of 5% in emotion accuracy.

We also include the results of distinct-1 and distinct-2 to demonstrate that the responses generated are varied and diversified. From assessment, it is clear that the proposed system is also competent enough to make the response diverse and interactive, along with producing emotional and persona-guided responses. By including the sentiment information in the contextual history, we see that there is improvement in the proposed framework as it facilitates in generating the correct emotional responses in accordance to the sentiment of the speaker.

We also perform an ablation study on the proposed framework to understand the importance of the emotion, persona and sentiment information in enhancing the performance of the overall framework. It is evident that the proposed framework compared to *Trans* and *Trans + E + P* models performs better as it also includes the sentiment of the previous utterances proving the significance of all the three components in generating better and interactive responses.

Human Evaluation Results: The manual evaluation, the results of which are recorded in Table 4, is carried out for a more comprehensive analysis of our proposed system. From the table, it is clear that the proposed system provides better performance with regards to all the specified metrics than

the existing approaches. As fluency calculates the grammatical accuracy of the response generated, it can, therefore, be assumed that the proposed model generates fluent responses. The final model having the highest scores in the case of fluency, as opposed to the baseline system, proves that the responses are grammatically correct and complete. Similarly, in the case of emotional content in the responses, we see that the frameworks having the emotion information seem to generate empathetic responses instead of the basic Seq2Seq and HRED frameworks. With an improvement of 6%, the proposed network surpasses the emotion score of the *HRED+E+P* model.

We also compute the ability of the models to maintain a consistent persona while generating the responses. From the manual evaluation results presented in the table, we can see that the *HRED+E+P* and *Seq2Seq+E+P* models show significant improvement from the typical HRED and Seq2Seq model in inducing the persona information while generating the responses. There is an enhancement in the proposed system from all of the other baseline systems in the case of the persona consistency metric. Also, the sentiment coherence score of the proposed framework is higher in comparison to the models without the sentiment information marking the importance of sentiment in the overall framework.

Through human assessment, it can, therefore, be inferred that the proposed system is capable of producing empathetic responses and has the capacity to retain a particular persona and respond with the correct sentiment.

Case Study and Discussion: In Table 5, we provide two examples and their corresponding generated responses by the different models. For both the examples, it is evident that the basic Seq2Seq and HRED frameworks generate short and non-emotional responses that do not increase user engagement. On the contrary, the baseline models having the knowledge of both persona and emotion generate empathetic and personalized responses. Moreover, the responses generated by our proposed framework are not only fluent but also are engaging, diverse, personalized, and emotionally appropriate to the conversational history and the sentiment.

We came across through some of the errors made by the baseline and proposed frameworks after performing a detailed comparative analysis of the generated responses. Some of the commonly occurring errors are: **(i) Extra information:** There are a few instances where the information in the input is found to be replicated, and extra words added, in both the baselines and the proposed system providing extra information. For example, Gold: *if I have time for cooking and repairing houses*; Predicted: *if I have time time hunting, cooking...* **(ii) Persona Discrepancy:** For certain instances, the responses generated by the proposed architecture are incompatible with the personality information and lack the precise details present in the speaker's persona texts. For example, Persona information: *I have 3 lovely kids and enjoy playing with them.* Predicted response: *I hate kids find them very annoying.*

6 Conclusion and Future Work

Grounding conversations based on the user's persona and emotions is an exciting research direction that can contribute to building natural, engaging and social conversational agents. Our current work presents one of the first examples of an empathetic, personalized dialogue generation for building a robust social chatbot using the sentiment information of the speaker in the ongoing conversation. We trained a novel transformer framework capable of generating responses that is sensitive to the emotions of the speaker and in accordance with their persona information and sentiment. Specifically, the experimental analysis on the PersonaChat dataset shows that the responses having both the emotional quotient and persona knowledge in the responses help build interactive and engaging conversations.

Our future work would focus on extending the architectural framework for improving the performance of the generation. In addition, we would also investigate other factors such as politeness, diversity in responses for creating a comprehensive social chatbot.

7 Ethical Declarations

All the resources used in this paper are publicly available. The dataset used in this paper is used only for the purpose of academic research. There is nothing to disclose that warrant the ethical issues.

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 PhD 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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Elia Bruni and Raquel Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288.

Stanley Chen, Douglas H. Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 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)*, pages 3734–3743.

Mauajama Firdaus, Naveen Thangavelu, Asif Ekba, and Pushpak Bhattacharyya. 2020. Persona aware response generation with emotions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 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*, pages 634–642.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 249–256.

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.

Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 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)*, pages 49–54.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Lee Kezar. 2018. Mixed feelings: Natural text generation with variable, coexistent affective categories. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018,Student Research Workshop*, pages 141–145.

Jingyuan Li and Xiao Sun. 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*, pages 678–683.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Jiwei Li, Will Monroe, Alan Ritter, Daniel Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1192–1202.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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*, pages 986–995.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2019. Caire: An end-to-end empathetic chatbot. *arXiv preprint arXiv:1907.12108*.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. 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 6794–6801.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *EMNLP*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. 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*, pages 5454–5459.

Bilyana Martinovski and David Traum. 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*, pages 11–16.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological review*, 50(4):370.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2775–2779.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. Association for Computational Linguistics.

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.

Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 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.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *IJCAI*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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*, pages 5370–5381.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81.

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.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5190–5196.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. *AAAI*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 53–65.

Xianchao Wu, Ander Martinez, and Momo Klyen. 2018. Dialog generation using multi-turn reasoning neural networks. 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)*, volume 1, pages 2049–2059.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949.

Zhen Xu, Bingquan Liu, Baoxun Wang, Cheng-Jie Sun, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via gan with an approximate embedding layer. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 617–626.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. *NeurIPS*.

Jiayi Zhang, Chongyang Tao, Zhenjing Xu, Qiaojing Xie, Wei Chen, and Rui Yan. 2019. Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–444.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *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 2204–2213.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 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), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *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 1128–1137.

Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. 2019. Retrieval-enhanced adversarial training for neural response generation. *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 3763–3773.