Multi-step Prompting for Few-shot Emotion-Grounded Conversations

Mauajama Firdaus[**]
University of Alberta
Alberta, Canada
mauzama.03@gmail.com

Gopendra Vikram Singh[**]
IIT Patna
Bihar, India
gopendra.99@gmail.com

Asif Ekbal
IIT Patna
Bihar, India
asif@iitp.ac.in

Pushpak Bhattacharyya
IIT Bombay
Maharashtra, India
pushpakbh@gmail.com

ABSTRACT

Conversational systems have shown immense growth in their ability to communicate like humans. With the emergence of large pre-trained language models (PLMs) the ability to provide informative responses have improved significantly. Despite the success of PLMs, the ability to identify and generate engaging and empathetic responses is largely dependent on labelled-data. In this work, we design a prompting approach that identifies the emotion of a given utterance and uses the emotion information for generating the appropriate responses for conversational systems. We propose a two-step prompting method that first recognises the emotion in the dialogue utterance and in the second-step uses the predicted emotion to prompt the PLM to generate the corresponding empathetic response in a few-shot setting. Experimental results on three publicly available datasets show that our proposed approach outperforms the state-of-the-art approaches for both automatic and manual evaluation.

CCS CONCEPTS

• **Natural Language Processing** → **Dialogue Generation**; • **Dialogues** → *Text,Emotion*; • **Deep Learning** → Prompting, Large Language Models.

KEYWORDS

empathy, LLMs, prompting, multi-step, few-shot

ACM Reference Format:
Mauajama Firdaus[**], Gopendra Vikram Singh[**], Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multi-step Prompting for Few-shot Emotion-Grounded Conversations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3583780.3615265

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615265

1 INTRODUCTION

Conversational systems have been known to assist humans in their day-to-day activities [50, 62, 63]. The emergence of Large Language Models (LLMs) [2, 8, 44, 45, 61] has had a significant impact on conversational systems, such as chatbots [40, 53, 60] and virtual assistants. With pre-trained models like GPT-3 [2], developers can leverage existing knowledge and create conversational systems that require less manual programming. Overall, the emergence of LLMs has greatly improved the capabilities and effectiveness of conversational systems, enabling them to provide more natural and contextually relevant interactions, as well as making them more accessible and easier to develop.

In recent years, there has been a growing interest in developing algorithms and models that can understand and generate emotional responses in human-machine interactions [15, 16, 20, 47]. One of the main challenges in emotion classification is that emotions are complex and multidimensional, and can be expressed in various ways, such as through language, facial expressions, and physiological signals [19, 25, 42, 54, 55]. While emotional response generation [14, 16, 17, 21, 34, 59, 66] aims at generating responses that are not only contextually relevant but also emotionally appropriate. This requires models to understand the emotional state of the user and generate responses that match their emotional state or change it if needed. Chatbots and virtual assistants that can detect and respond to customers' emotions can help build rapport and trust, and ultimately improve customer loyalty and retention [18, 22, 36–38, 41, 58].

Few-shot emotion classification [24] refers to the ability of NLP models to recognize emotions from text, with limited training examples. Few-shot dialogue generation [3, 6, 57] extends the capability to situations where the model has only a few examples of responses to learn from. Few-shot classification/generation can be useful in scenarios where training data is limited or where the model needs to adapt quickly to new or changing emotional expressions. Overall, few-shot emotion classification and emotional dialogue generation represent exciting advancements in NLP and conversational AI, with promising applications in various fields. While there are still challenges to be addressed, these techniques have the potential to improve the emotional intelligence of machines, enabling them to better understand and respond to human emotions.

[**] The first two authors contributed equally to this work and are jointly first authors.

Recent advances in prompting [11, 31, 51] have focused on enabling language models to perform complex tasks with minimal or no training examples. Techniques like few-shot [23, 33, 49] and zero-shot learning [39, 48, 65] allow models to generalize from a few examples or even perform tasks they haven't been explicitly trained on. This has significant implications for applications requiring rapid adaptation to new domains or tasks.

In our current work, we propose the task of jointly identifying the emotions in conversation followed by the empathetic response generation in a few-shot setting. For this, we design a two-step promoting approach where we first identify the emotion of the given sentence and use the emotion as knowledge for the next response generation. The major contributions of our work are threefold: (i) we propose the task of jointly performing emotion recognition and empathetic response generation; (ii) we design a novel multi-step prompting approach for building empathetic end-to-end dialog system; (iii) experimental analysis on three dialogue datasets show that our proposed approach performs better than the existing methods.

2 RELATED WORK

There have been several recent advancements in textual emotion classification in conversations, focusing on understanding and recognizing emotions expressed within conversational contexts [1, 26, 29, 30, 46]. Kim and Vossen [26] employs a straightforward approach to capture speaker information and contextual cues in conversations. Bao et al. [1] introduces a unique approach to speaker modeling that considers both intra- and inter-speaker dependencies dynamically. Additionally, they present a Speaker-Guided Encoder-Decoder (SGED) framework for Emotion Recognition in Conversations (ERC) that effectively utilizes speaker information during emotion decoding. Recently, the authors in [4] proposed the pre-finetuning of speech models on challenging tasks to transfer knowledge to downstream few-shot classification objectives.

Recent advances in empathetic response generation have significantly improved the ability of conversational AI models to generate emotionally appropriate and empathetic responses [12, 32, 43]. Cheng et al. [7] introduces a novel system called MultiESC that aims to tackle these challenges. To enhance strategy planning, it draws inspiration from the A* search algorithm and proposes lookahead heuristics. The authors [27] utilize external knowledge, such as commonsense knowledge and emotional lexical knowledge, as a means to explicitly comprehend and convey emotions in empathetic dialogue generation.

3 APPROACH

Our proposed multi-step emotion-grounded (MSEG) dialogue prompting framework is depicted in Figure 1. The MSEG prompting approach consists of an emotion predictor and a response generator, both using the same pre-trained LM. The emotion predictor produces the emotion label to the input utterance having the dialogue history as context. While the response generator generates empathetic and interactive responses based on the dialogue context and the predicted emotion.

Emotion Predictor. To avoid the dependency on a large-scale labeled emotion dataset, we propose a prompt-based emotion classification approach, which uses a relatively small set of in-context exemplars (about 10 samples for each emotion) and a pre-trained LM to predict the appropriate emotion labels as shown in Figure 1.

Sample Selection. We hypothesize that choosing relevant samples as prompts is the key to achieving correct emotion labels for the given utterance. Hence, we adopt a query-based sample-selection method that aims to search related samples from **D** based on the input utterance (i.e., query (**q**)). To ensure that the selected examples are appropriate to the query, we employ a pre-trained sentence encoder (SE) [10] to obtain the representation for the query and each data sample (d_i) in **D**. Then, we calculate the similarity between the query and each sample using the dot product of their representation as follows:

$$Sim(q, d_i) = SE(e + h)^T .SE(e_i + H_i), \qquad (1)$$

where the input of the SE is a concatenation of the emotion and dialogue utterance pair in the conversational history. Finally, we select n samples that have the highest similarity scores to q. This selection process can be performed effectively since the database is relatively small.

Emotion Classification. Inspired by the few-shot approach in Brown et al. [2], feeding the pre-trained LM with correct and intuitive prompts can allow it to generate relevant emotion labels. Specifically, the prompt for the i^{th} sample ($prompt_i$, $i \in [1, n]$) is $h_i^* \Rightarrow e_i$[**], and the prompt for the current utterance ($prompt_{curr}$) is $t(U^*) \Rightarrow$, here we use the ⇒ to guide the LM for generating the emotion label. Finally, we concatenate the constructed prompts using \n and feed them into the LM to generate the emotion label:

$$e' = LM(prompt_1\backslash n \cdots prompt_n\backslash nprompt_{curr}) \qquad (2)$$

where e' denotes the generated emotion label for the given input utterance.

Empathetic Response Generation. Finetuning an LM could lead to overfitting when the finetuning dataset is comparatively small. Also, one can access the gigantic LMs, like GPT-3 [2] and Megatron-Turing NLG 530B [56] using only APIs, finetuning them might not be a feasible solution. Therefore, we propose to circumvent the finetuning by prompting the pre-trained LM for the empathetic response generation, which requires a few dialogue examples. To generate emotional and engaging responses, we focus first on selecting the appropriate samples and then utilize them to effectively prompt the LM for response generation.

Sample Selection. One of the essential skills for empathetic response generation is to efficiently leverage the emotional information produced in the first stage, in order to make the responses empathetic. Considering that we can provide the LM with only a few dialogue samples, it could be difficult for it to learn how to generate a response based on the given emotion category unless there is a strong connection between the responses and the emotion labels that we provide. Concretely, for each example in the database, we calculate how similar they are to the given emotion label using the cosine similarity.

Response Generation. Aside from the ability to utilize the predicted emotion, another essential skill for the response model is to have the ability to chat based on the dialogue context. To equip our model with this skill, we focus on constructing intuitive prompts for the selected examples and feed them into the LM. For prompts

[**] For example, *(I hate being surrounded by such irritating audience.)* ⇒ *Anger*.

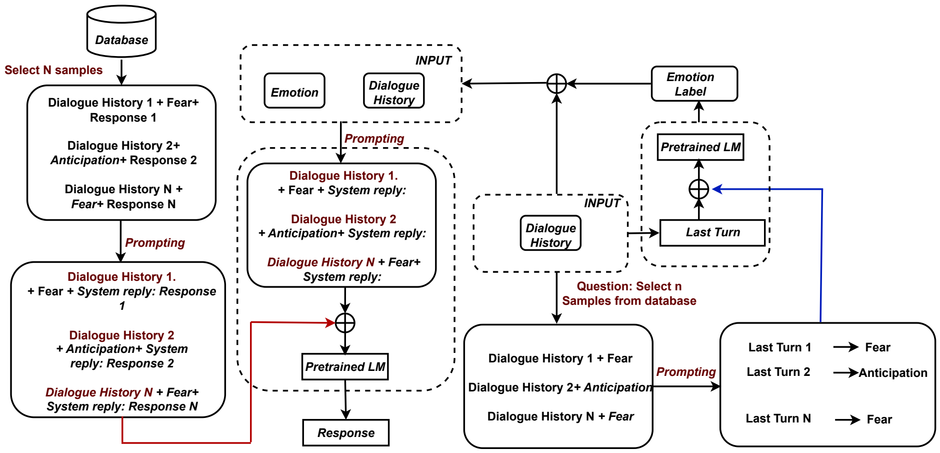
Database
Select N samples
Dialogue History 1 + Fear+ Response 1
Dialogue History 2+ Anticipation+ Response 2
Dialogue History N + Fear+ Response N
Prompting
Dialogue History 1. + Fear + System reply: Response 1
Dialogue History 2 + Anticipation+ System reply: Response 2
Dialogue History N + Fear+ System reply: Response N
INPUT
Emotion
Dialogue History
Prompting
Dialogue History 1. + Fear + System reply:
Dialogue History 2 + Anticipation+ System reply:
Dialogue History N + Fear+ System reply:
Pretrained LM
Response
Emotion Label
Pretrained LM
INPUT
Dialogue History
Last Turn
Question: Select n Samples from database
Dialogue History 1 + Fear
Dialogue History 2+ Anticipation
Dialogue History N + Fear
Prompting
Last Turn 1 Fear
Last Turn 2 Anticipation
Last Turn N Fear

Figure 1: Architectural diagram of our proposed MSEG framework

from the selected examples, we use the speaker information (such as *A* and *B* for the DailyDialog dataset) to connect different turns in the conversational history, and *We know that the emotion is:* and the *Speaker replies:* are used to inculcate the emotion knowledge and response, respectively. For prompts from the current conversation (i.e., inputs), we follow the same template except that we keep the response empty for the pre-trained LM to generate. After the prompt construction, we concatenate the prompts for selected samples and the inputs using \n and then feed them into the pre-trained LM to generate the response.

4 EXPERIMENTAL SETUP

Datasets. We evaluate our model using three emotionally-labeled datasets: Dailydialog [28], EmotionLines [5] and EmoWOZ [13]. DailyDialog consists of the emotion labels *anger, fear, disgust, happiness, sadness, surprise, other.* It comprises 11118 dialogues in train while 1000 dialogues each in validation and test set. EmotionLines is annotated with Ekman's six universal emotions *(Joy, Sadness, Fear, Anger, Surprise, and Disgust)* Finally, 250 dialogues were sampled randomly from each of these groups, resulting in the final dataset of 1,000 dialogues. EmoWOZ [13] consists of the emotion categories *neutral, fearful, dissatisfied, apologetic, abusive, excited, satisfied.* It consists of 8825 dialogues in train while 1041 dialogues each in validation and test set.

Implementation Details. The LMs used for our MSEG model, and baselines are GPT-style [2] models and are pre-trained using the toolkit in Shoeybi et al. [52]. PPLM uses dialoGPT-medium, which has 355 million parameters(335m). The LM in FCM has 357m parameters. To test how different model sizes affect the performance, we evaluate our methods with 126m, 357m, 1.3 billion (1.3b), and 530 billion (530b) parameters LMs. For the sample selection, we choose 15 samples for prompting in the emotion prediction module and 20 samples for the prompting in response generation module.

Evaluation Metrics. For emotion prediction, we use the F1 score while for response generation we calculate the BLEU score, Rouge-L, and emotion accuracy (EA) in the generated response, respectively for automatic evaluation. For manual evaluation, we randomly select two hundred responses from each dataset. Five well-defined metrics are used: fluency, knowledge consistency, context relevance/coherence, informativeness, and engagingness. Fluency is assessed to gauge the naturalness and coherence of the generated response. Emotion consistency is employed to evaluate whether the generated response utilizes appropriate emotion. Context relevance/coherence is employed to assess the degree to which the generated response aligns with the given situation or discussion. Informativeness is utilized to ascertain the level of information provided by the generated response. Engagingness is employed to verify if the generated responses align with the user's conversational objectives. These metrics are scored on a range of 0 to 5, with 0 indicating errors and 5 denoting the highest quality response.

Baselines. For emotion prediction, we consider the existing few-shot baseline ProtoSeq [24]. While for response generation, we consider **PPLM** [9] that denotes the plug and play language model. We choose it as a baseline because our MSEG can be considered as using emotion to control the LM to generate responses, and PPLM, which does not need finetuning either, can be also used to control LMs for emotion-guided generation. We follow Madotto et al. [35] and use dialoGPT [64] for PPLM to enable response generation. Another baseline, **FCM** denotes the finetuning-based conversational model. We use the training dataset of DailyDialog to finetune the LM. This baseline has the same pipeline as that of our MSEG. Instead of doing prompting, it uses the ground-truth emotion labels to guide the FCM for empathetic response generation.

5 RESULTS

Automatic Evaluation Results. In Table 1, we present the results of automatic evaluation for both the tasks i.e., emotion prediction and emotional response generation on all the three emotion-annotated datasets. For the emotion prediction task, in comparison to ProtoSeq [24] our proposed MSEG framework performs significantly better for all the datasets. In addition, we see that the MSEG framework having 530b parameters performs the best as opposed to the other frameworks having lesser parameters.

For the response generation, we compare with two baselines i.e, PPLM and FCM. The MSEG framework, as obvious with the most

Models		DailyDialog				EmotionLines				EmoWOZ			
		Emotion Prediction	Response Generation			Emotion Prediction	Response Generation			Emotion Prediction	Response Generation		
		E-F1	BLEU	Rouge-L	EA	E-F1	BLEU	Rouge-L	EA	E-F1	BLEU	Rouge-L	EA
Baselines	*ProtoSeq* [24]	61.71	-	-	-	65.69	-	-	-	63.71	-	-	-
	PPLM [9]	-	4.67	9.05	38.71	-	5.01	10.98	39.93	-	4.10	8.32	37.41
	FCM	-	6.18	10.23	40.56	-	6.91	11.69	42.35	-	5.67	9.76	39.11
Proposed Approach	*MSEG-126m*	64.83	9.65	14.81	45.29	68.73	10.87	15.61	47.15	66.83	9.01	13.47	44.01
	MSEG-357m	67.71	10.59	15.75	47.19	69.35	11.87	16.91	48.65	68.61	9.88	14.31	46.07
	MSEG-1.3b	68.26	11.88	16.49	47.91	71.43	13.11	17.84	49.23	69.16	10.53	15.16	46.75
	MSEG-530b	71.64	13.93	18.68	50.37	74.43	15.24	19.71	52.61	73.01	12.76	17.17	49.67
Ablation Study	*MSEG w/o EP*	-	12.69	16.87	48.61	-	13.32	17.92	49.54	-	11.42	15.39	47.43
	MSEG w/o RG	69.85	-	-	-	72.29	-	-	-	71.91	-	-	-

Table 1: Automatic Evaluation Results on three different datasets for both Emotion Prediction and Emotional Response Generation; here EA: Emotion Accuracy

Models		Daily Dialog					EmotionLines					EmoWOZ				
		F	EC	Cr	Info	Eng	F	EC	Cr	Info	Eng	F	EC	Cr	Info	Eng
Baselines	PPLM	1.03	0.69	1.08	1.12	0.98	1.29	0.83	1.11	1.23	1.05	1.31	0.97	1.19	1.44	1.27
	FCM	1.29	1.01	1.31	1.31	1.20	1.55	1.41	1.30	1.57	1.39	1.53	1.33	1.42	1.57	1.71
Proposed Approach	MSEG-126m	2.81	2.78	2.35	2.41	2.55	2.80	2.85	2.99	2.89	2.98	2.51	2.21	2.89	2.34	2.55
	MSEG-357m	3.11	3.01	2.51	2.79	2.86	3.03	2.79	3.27	3.09	3.21	2.98	2.49	3.11	2.61	2.79
	MSEG-1.3b	3.30	3.23	2.77	3.15	3.39	3.24	3.15	3.31	3.53	3.89	3.24	2.78	3.32	2.99	3.03
	MSEG-530b	3.63	3.52	3.17	3.45	3.98	4.04	3.75	3.95	4.07	4.11	3.51	3.28	3.13	3.31	3.41

Table 2: Manual evaluation. Here, F, EC, Cr, Info, and Eng refer to fluency, emotion consistency, context relevance/coherence, informativeness, and engaginess respectively.

Conversation: DailyDialog *- That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.* ***Emotion from MSEG (530b):*** *- Disgust* ***Response from MSEG (530b):*** *- Yes, I will definitely try to breathe and work on my issues.*
Conversation: EmotionLines *- You Liked it? Oh my God! You really like it.* ***Emotion from MSEG (530b):*** *- Surprise* ***Response from MSEG (530b):*** *- Oh yeah!!! It was amazing I enjoyed every part of the evening like you.*
Conversation: EmoWOZ *- I am afraid that's all I've got.* ***Emotion from MSEG (530b):*** *- Apologetic* ***Response from MSEG (530b):*** *- Don't be sorry, please could you book me from Petersborough to Cambridge on Saturday.*

Table 3: Case studies on different datasets for our best performing MSEG model

number of parameters outperforms all the other variants. In comparison to the PPLM approach, we see a gain of about 5 points in BLEU score for the MSEG framework with only 126m parameters. While for the MSEG having 530b parameters the performance improvement is more than 9 and 7 points for PPLM and FCM frameworks respectively for the DailyDialog dataset. Similar performance gain is seen for the EmotionLines as well as the EmoWOZ datasets. For the Rouge-L metric, both the EmotionLines and EmoWOZ datasets achieves an increase of approximately 8 and 9 points compared to the fine-tuned model (FCM).

The sole purpose of our work is to build an empathetic conversational model, therefore we also evaluate the emotion accuracy in the generated responses for all the different datasets. From the results we see that the generated responses achieve a considerable accuracy for emotions ensuring an end-to-end approach for the creation of emotional dialogue system. We also perform the ablation study of our MSEG framework. The MSEG w/o RG uses the 530b parameters and we see that there is a decrease in performance for the emotion prediction task of approximately 2 points all the different datasets. Similarly for the MSEG w/o EP there is a significant drop in all the three metrics such as BLEU, Rouge-L and emotion accuracy for all the datasets. The ablation study proves that both the steps is equally significant and enhances the performance of the overall model.

Manual Evaluation Results. In Table 2, we provide the results of manual evaluation to verify the quality of the generated response. From the table, it is evident that the responses generated by the proposed framework having different parameters are more fluent compared to the existing PPLM and FCM models for all the three datasets. Also, the informativeness and coherence metric shows a jump in scores for DailyDialog, EmotionLines and EmoWOZ indicating that the generated responses are not only gramatically correct but also are capable of retaining the information and is coherent to the ongoing dialogue context. The emotional consistency of the proposed approach compared to fine-tuned model increases by an absolute 2.5, 2 and 1.9 points for DailyDialog, EmotionLines and EmoWOZ datasets respectively. In addition, the engaginess metric increases signifying the responses are interactive and engaging.

In Table 3, we provide few case studies on our best performing MSEG framework on all the three datasets. From the Table it is evident that the MSEG framework is capable of identifying the correct emotions of the given utterance and also generate an engaging empathetic and coherent response.

6 CONCLUSION

An empathetic conversational system refers to a chatbot or conversational AI system that is designed to understand and respond to human emotions with empathy and sensitivity. In our current work, we devise a multi-step prompting approach for identifying the emotions in the responses and concurrently use the identified emotions for the generation of the next response. We evaluated our proposed MSEG framework on three different dialogue datasets such as DailyDialog, EmotionLines and EmoWOZ datasets. Both quantitative and qualitative analysis shows that the proposed MSEG framework performs significantly better than the existing baselines.

REFERENCES

[1] Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. *arXiv preprint arXiv:2206.03173* (2022).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text generation. *arXiv preprint arXiv:2107.03176* (2021).

[4] Maximillian Chen and Zhou Yu. 2023. Pre-Finetuning for Few-Shot Emotional Speech Recognition. *arXiv preprint arXiv:2302.12921* (2023).

[5] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* (2018).

[6] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2019. Few-shot NLG with pre-trained language model. *arXiv preprint arXiv:1904.09521* (2019).

[7] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. *arXiv preprint arXiv:2210.04242* (2022).

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[9] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* (2019).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998* (2021).

[12] Chengzhang Dong, Chenyang Huang, Osmar Zaïane, and Lili Mou. 2021. Simulated annealing for emotional dialogue systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2984–2988.

[13] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. 2021. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. *arXiv preprint arXiv:2109.04919* (2021).

[14] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1555–1566.

[15] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. 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.

[16] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2021. More the merrier: Towards multi-emotion and intensity controllable response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12821–12829.

[17] Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021. SEPRG: sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*. 353–363.

[18] Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems* (2022).

[19] Mauajama Firdaus, Gopendra Vikram Singh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Affect-GCN: a multimodal graph convolutional network for multi-emotion with intensity recognition and sentiment analysis in dialogues. *Multimedia Tools and Applications* (2023), 1–22.

[20] Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2022. Sentiment guided aspect conditioned dialogue generation in a multimodal system. In *European Conference on Information Retrieval*. Springer, 199–214.

[21] Mauajama Firdaus, Naveen Thangavelu, Asif Ekbal, and Pushpak Bhattacharyya. 2022. I enjoy writing and playing, do you: A Personalized and Emotion Grounded Dialogue Agent using Generative Adversarial Network. *IEEE Transactions on Affective Computing* (2022).

[22] Hitesh Golchha, **Mauajama Firdaus**, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 851–860.

[23] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332* (2021).

[24] Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-shot emotion recognition in conversation with sequential prototypical networks. *arXiv preprint arXiv:2109.09366* (2021).

[25] Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Seq2Emo for Multi-label Emotion Classification Based on Latent Variable Chains Transformation. *arXiv preprint arXiv:1911.02147* (2019).

[26] Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009* (2021).

[27] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10993–11001.

[28] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[29] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504* (2022).

[30] Chen Liang, Chong Yang, Jing Xu, Juyang Huang, Yongliang Wang, and Yang Dong. 2021. S+ page: A speaker and position-aware graph neural network model for emotion recognition in conversation. *arXiv preprint arXiv:2112.12389* (2021).

[31] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[32] Yiren Liu and Halil Kilicoglu. 2023. Commonsense-Aware Prompting for Controllable Empathetic Dialogue Generation. *arXiv preprint arXiv:2302.01441* (2023).

[33] Robert L Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353* (2021).

[34] Avinash Madasu, Mauajama Firdaus, and Asif Eqbal. 2022. A Unified Framework for Emotion Identification and Generation in Dialogues. *arXiv preprint arXiv:2205.15513* (2022).

[35] Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. *arXiv preprint arXiv:2010.04344* (2020).

[36] 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*. 11–16.

[37] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing* 494 (2022), 242–254.

[38] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. GenPADS: Reinforcing politeness in an end-to-end dialogue system. *Plos one* 18, 1 (2023), e0278323.

[39] Nihal V Nayak, Peilin Yu, and Stephen H Bach. 2022. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574* (2022).

[40] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309* (2022).

[41] 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 (2005), 231–245.

[42] Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2023. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications* 224 (2023), 120025.

[43] Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think Twice: A Human-like Two-stage Conversational Agent for Emotional Response Generation. *arXiv preprint arXiv:2301.04907* (2023).

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[46] Waleed Ragheb, Mehdi Mirzapour, Ali Delfardi, Hélène Jacquenet, and Lawrence Carbon. 2022. Emotional Speech Recognition with Pre-trained Deep Visual Models. *arXiv preprint arXiv:2204.03561* (2022).

[47] Azlaan Mustafa Samad, Kshitij Mishra, **Mauajama Firdaus**, and Asif Ekbal. 2022. Empathetic Persuasion: Reinforcing Empathy and Persuasiveness in Dialogue Systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 844–856.

[48] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).

[49] Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. *arXiv preprint arXiv:2010.13641* (2020).

[50] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers* (2015), 1577–1586.

[51] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).

[52] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).

[53] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188* (2022).

[54] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInt-Trans: A Multimodal Transformer for Identifying Emotions and Intents in Social Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 290–300.

[55] Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. *arXiv preprint arXiv:2205.13908* (2022).

[56] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990* (2022).

[57] Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2019. Learning to customize model structures for few-shot dialogue generation tasks. *arXiv preprint arXiv:1910.14326* (2019).

[58] **Mauajama Firdaus**, Asif Ekbal, and Pushpak Bhattacharyya. 2022. PoliSe: Reinforcing Politeness using User Sentiment for Customer Care Response Generation. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. 6165–6175.

[59] **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)*. IEEE, 1–8.

[60] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[62] Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869* (2015).

[63] 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* (2018), 2204–2213.

[64] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).

[65] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670* (2021).

[66] 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), New Orleans, Louisiana, USA, February 2-7, 2018*. 730–739.