

ASEM: Enhancing Empathy in Chatbot through Attention-based Sentiment and Emotion Modeling

Omama Hamad¹, Ali Hamdi², Khaled Shaban¹

¹Qatar University, Department of Computer Science and Engineering, Qatar

²MSA University, Department of Computer Science, Egypt

{omama.hamad, khaled.shaban}@qu.edu.qa, ahamdi@msa.edu.eg

Abstract

Effective feature representations play a critical role in enhancing the performance of text generation models that rely on deep neural networks. However, current approaches suffer from several drawbacks, such as the inability to capture the deep semantics of language and sensitivity to minor input variations, resulting in significant changes in the generated text. In this paper, we present a novel solution to these challenges by employing a mixture of experts, multiple encoders, to offer distinct perspectives on the emotional state of the user's utterance while simultaneously enhancing performance. We propose an end-to-end model architecture called ASEM that performs emotion analysis on top of sentiment analysis for open-domain chatbots, enabling the generation of empathetic responses that are fluent and relevant. In contrast to traditional attention mechanisms, the proposed model employs a specialized attention strategy that uniquely zeroes in on sentiment and emotion nuances within the user's utterance. This ensures the generation of context-rich representations tailored to the underlying emotional tone and sentiment intricacies of the text. Our approach outperforms existing methods for generating empathetic embeddings, providing empathetic and diverse responses. The performance of our proposed model significantly exceeds that of existing models, enhancing emotion detection accuracy by 6.2% and lexical diversity by 1.4%. ASEM code is released at <https://github.com/MIRAH-Official/Empathetic-Chatbot-ASEM.git>.

Keywords: dialogue, attention, chatbot

1. Introduction

The ability of deep learning models to generate natural language is highly dependent on the quality of the feature representations, a.k.a., embeddings, that capture the underlying patterns in text. Thus, representation learning is crucial for various natural language processing tasks, including text generation, which mimics a human skill. However, this task can be challenging due to colloquial dialects and the diverse ways emotions are expressed, blurring the boundaries between distinct emotional states. Additionally, certain emotions, such as surprise and fear, exhibit close relationships as evidenced by psychological studies, which can lead to confusion in the model's understanding (Noordewier and Breugelmans (2013)). Consequently, developing high-quality text generation models requires complex architecture, extensive data resources, and an understanding of linguistic nuances. Recent advancements have been made in utilizing different techniques to enhance and learn representations (Wu et al. (2021); El Boukkouri et al. (2022)).

In this work, we propose an effective attention-based model for empathetic text generation in open-domain chatbots, aiming to engage and interact with users in a natural, human-like manner. Furthermore, the model is trained with empathetic conversations to create a more immersive user experience and align with their emotional states. Previous stud-

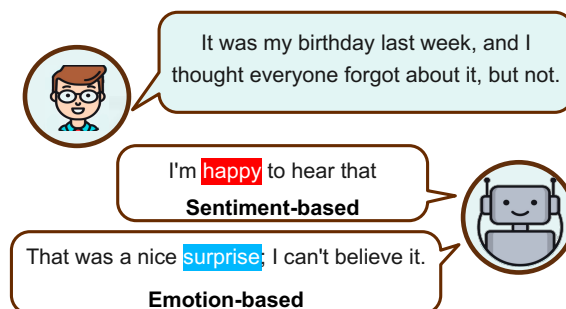


Figure 1: Emotional vs Sentimental Response.

ies have employed sentiment or emotion modelling to address empathetic conversational abilities (Zarnis et al. (2021); Majumder et al. (2020)). Sentiment refers to the overall tone or mood of a user's utterance, while emotion represents specific feelings or mental states, encompassing over 30 emotions. Emotions are often described along dimensions like valence (positive or negative), arousal (high or low), and dominance (high or low) (Plutchik (1980)). Although sentiment analysis offers insights into the overall tone, it may not capture the full spectrum of emotions a user experiences. On the other hand, emotion analysis helps identify specific emotions, such as joy, fear, or surprise.

Consequently, relying solely on sentiment analysis may cause confusion in the generated embeddings when texts with the same sentiment belong

to different emotions. For instance, statements like "That was a nice surprise" and "I'm happy to hear that" may express similar sentiment (positive) but convey different emotions (surprise and happiness) as depicted in Figure 1. This highlights the importance of emotion analysis in understanding the user's feelings and generating appropriate responses. Hence, we incorporate emotion analysis on top of sentiment analysis using a mixture of experts to enable the empathetic chatbot to comprehend the user's emotional state more comprehensively. This approach facilitates the generation of empathetic responses tailored to the user's specific emotional needs. For example, if a user expresses sadness in their message, sentiment analysis may detect a negative sentiment, but emotion analysis can identify the specific emotion of sadness. Consequently, the chatbot can then generate a response that acknowledges the user's sadness and provides words of comfort or support. Introducing a pioneering attention mechanism, the proposed model uniquely discerns and prioritizes sentiment- and emotion-laden aspects of a user's utterance, ensuring the generation of representations that deeply capture the intricate contextual nuances within the text. The proposed approach enables the chatbot to provide a score indicating how well the generated embedding matches the user's sentiment. Additionally, the magnitude of each sentiment serves as an indicator of how strongly each expert should contribute to the total score. For example, as depicted in Figure 2, a statement may exhibit a combination of positive, negative, and neutral sentiments.

Our contributions can be summarized as follows:

- **Sentiment and Emotion Experts:** We propose a comprehensive model for open-domain chatbots that can recognize the sentiment and emotion in user statements, enabling the generation of empathetic responses tailored to the user context.
- **The use of a specialized attention strategy:** We compute the score of each expert based on the proposed attention mechanism, improving the model's ability to generate empathetic responses. This improvement was demonstrated through both automatic and human evaluations.
- **Standardization of emotion categories:** We establish a standardized set of emotion categories across two datasets, enhancing consistency and comparability in emotion analysis.

Overall, our proposed ASEM model enhances the chatbot's empathetic capabilities, providing more accurate sentiment and emotion recognition, and generating contextually relevant responses that resonate with users' emotional states.

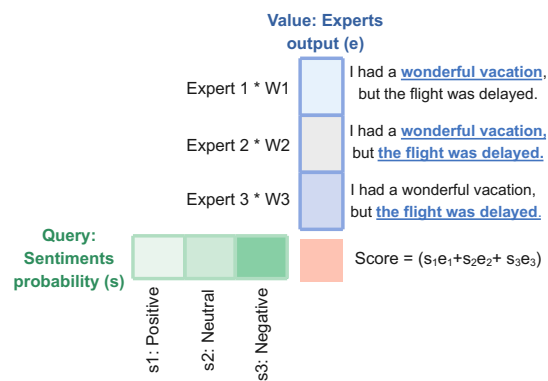


Figure 2: Attention Weights.

2. Related Work

Open-domain chatbots are AI-based systems designed to engage and entertain humans in conversational interactions. These chatbots require various skills to effectively emulate human conversations, and empathy is considered one of the most important skills. Empathy involves understanding and being attuned to the emotional states of users. Psychological studies have identified different methods through which humans express empathy [Baron-Cohen \(2006\)](#); [Low \(2012\)](#); [Song \(2015\)](#). These methods include 1) expressing care and concern, 2) acknowledging the user's feelings to avoid closed-off responses such as 'ok' or 'I see', 3) asking questions that encourage sharing of feelings and 4) demonstrating emotion-supporting behaviour. Each of these cases is illustrated in Table 1. However, existing research that relies on human evaluations of empathetic scenarios often does not specify the level of empathy involved, ranging from weak to robust. Therefore, we propose a framework based on these four cases for human evaluation of empathetic responses generated by the chatbot.

Recent research has explored various perspectives in this domain. For example, some studies focused on understanding sentiments in user inputs to generate responses with a sentimental tone [Zarannis et al. \(2021\)](#). [Firdaus et al. \(2021\)](#) proposed a transformer-based model for generating emotional responses aligned with the user's persona and conditioned on a specific sentiment. However, relying solely on sentiment may not always yield the desired emotion, as emotional analysis is more comprehensive [Kumar et al. \(2019\)](#). Multitask architectures have been employed to jointly learn and classify sentiments and emotions, utilizing attention mechanisms to incorporate information from various sources [Akhtar et al. \(2019\)](#). Another recent study introduced a multi-party empathetic dialogue generation task, employing a dynamic graph network to incorporate temporal information and blend dynamic emotions and static sensibilities from differ-

ent parties [Zhu et al. \(2022\)](#). Temporal and spatial information can be crucial for empathetic text generation, enabling the understanding of local entity relationships and interactions. Additionally, there have been approaches based on fine-grained or coarse-grained emotion analysis [Lin et al. \(2019\)](#); [Naous et al. \(2020\)](#). For example, open-domain chatbots like Xiaolce have demonstrated the ability to understand users’ emotional needs, but their generated responses may not always be accurate and rely on external knowledge [Zhou et al. \(2020\)](#). Other end-to-end approaches have encountered challenges in generating irrelevant or low-diversity responses [Zhang et al. \(2020\)](#).

In summary, while existing research on open-domain chatbots has made significant strides in understanding and generating empathetic responses, a clear gap remains in the application and evaluation of empathy in these interactions. Current studies often do not define the varying levels of empathetic engagement, ranging from basic acknowledgment to deep emotional resonance, which are crucial for human-like conversational experiences. Moreover, the reliance on sentiment analysis alone has proven insufficient for capturing the full spectrum of human emotions and their context within conversations. Our proposed work aims to bridge this gap by introducing a structured and comprehensive approach to evaluating empathetic responses in chatbots. By focusing on a multi-dimensional evaluation that encompasses both the depth of empathy and the integration of emotion analysis on top of sentiment analysis to accurately acknowledge and understand empathy, we intend to enhance the empathetic capabilities of chatbots, making them more adept at navigating complex emotional landscapes in human interactions.

3. Methodology

3.1. Problem Formulation

In multi-turn conversational settings, the problem is formalized as exchanges between a user’s statement U and a chatbot’s response R , represented by the overall user context as $C = \{U_1, R_1, U_2, R_2, \dots\}$. The emotion of the user is expressed in each user statement, such as joy or surprise, and belongs to a sentiment class, which can be positive, negative or neutral, denoted as $E = \{e_j, e_l, \dots, e_n\}$ and $S = \{s_{pos}, s_{neg}, s_{neu}\}$, respectively. The user’s history C includes all previous exchanges and context, and excludes the most recent statement by the user. The objective of the model is to identify the sentiment and emotion of the user’s statement based on the current statement and the conversation history, and generate a relevant, coherent and empathetic response.

Context	I’m having to cancel my Labor Day plans since I have to work. I’m disappointed.
Responses	
Case 1	How do you feel now?
Case 2	I’m sorry to hear that. I realise that what you are going through is very difficult.
Case 3	Oh no, I’m so sorry to hear that. Is there anyone who can help you?
Case 4	I’m sorry to hear that. Is there anything I can do for you?

Table 1: Empathetic Cases.

3.2. Proposed Model

We propose a novel approach that sequentially tracks emotions and sentiments to create a more effective and nuanced chatbot. Our proposed method, illustrated in Figure 3, consists of two stages: learning empathetic feature representations and generating an empathetic response. The first stage includes a sentiment-tracker and an emotion tracker, while the second stage adopts the concept of a mixture of empathetic listeners from [Lin et al. \(2019\)](#) for generating an empathetic response. A key novelty in our approach is the mechanism to combine the outputs of these experts. Instead of directly pooling or averaging the experts’ outputs, our model computes the mean of all expert outputs and multiplies them with the sentiment probability distributions. This ensures that each sentiment-specific feature is weighted appropriately based on its relevance, producing a sentiment-specific context-rich feature representation. Furthermore, by adding the weighted feature representation with the general encoder’s output, the model not only retains the original semantic information but also infuses it with sentiment-aware nuances. This enriched representation is then used to predict emotion classes, thereby grounding the model’s attention mechanism in both semantic and emotional dimensions of the text.

3.2.1. Stage 1: Learning Empathetic Features Representation

General Encoder: The general encoder is responsible for generating feature representations of sequences G_H with corresponding probability distributions for different sentiment classes S_{att} . These probability distributions are used as attention weights for the next group of experts. The general encoder takes two inputs: the current user turn U and the previous conversation history C , both represented by GloVe embeddings. The encoder consists of multi-head attention and position-wise feed-forward layers.

To control the importance of the current turn compared to the conversation history U_w , a weighting

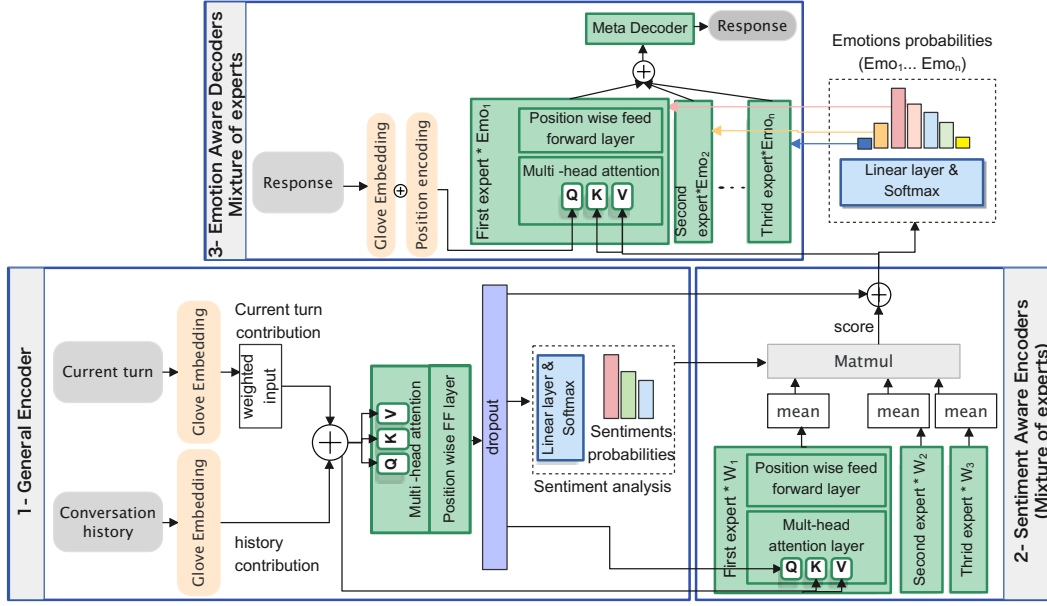


Figure 3: ASEM model for empathetic chatbot.

factor is applied to the embeddings of the current turn. This weighting factor allows the model to learn to focus more on recent information while still considering the context provided by the overall conversation history. The weighted representation of the current turn is then combined with the conversation history and fed into the encoder:

$$U_w = Emb(U) \times W \quad (1)$$

$$G_h = Enc([U_w, Emb(C)]) \quad (2)$$

Here, $G_h \in \mathbb{R}^{L \times d}$ represents the last hidden state of the general encoder, where L is the length of the conversation and d is the hidden dimension of the encoder.

To obtain the probability distribution S_{att} , a softmax function is applied to G_H :

$$S_{att} = Softmax(G_H) \quad (3)$$

The performance of the general encoder is optimized by calculating the loss using the cross-entropy function, which measures the difference between the predicted sentiment label p_s and the actual label y_s .

$$L_1 = -y_s \log(S_{att}) \quad (4)$$

Sentiment Aware Encoders (Mixture of Experts): Sentiment Aware Encoders (SAEs) are designed to generate sentiment-specific feature representations that capture the contextual nuances of the input text. Unlike conventional encoders that produce fixed representations, SAEs employ a set of encoders, each initialized with different weights and dedicated to modeling a specific sentiment.

The last hidden state of the general encoder, denoted as G_h , serves as query input to each encoder in the attention layer, while the concatenated input C is the key and value. The context of each input is captured by taking the mean of the output tensors of each expert, allowing us to combine the outputs and generate a more informative representation S_h of the input text. By performing a batch matrix multiplication between the sentiment probability distributions S_{att} and the output of the expert mean tensor, we obtain a sentiment-specific context-rich feature representation. This computation can be performed efficiently by computing the weighted sum for all examples in the batch in a single operation.

$$S_h = \bar{E}_0((G_h, Emb(C)) * W_i), \dots, \bar{E}_i \quad (5)$$

where \bar{E}_i represents the output of the i -th expert encoder and W_i represents the weight matrices for the i -th encoder.

To weight the output features S_h based on their relevance, the attention score W_{att} is calculated as the weighted sum of the outputs of the three experts k :

$$W_{att_{i,j,l}} = \sum_k S_{att_{i,j,k}} \times S_{h_{i,k,l}} \quad (6)$$

Here, the attention-weighted output $W_{att_{i,j,l}}$ for each data instance i in the batch, feature j , and hidden dimension l is obtained by summing over the product of attention scores $S_{att_{i,j,k}}$ and the corresponding expert outputs $S_{h_{i,k,l}}$ for all experts k . The attention scores serve as weights that determine the significance of each expert's contribution, effectively allowing the model to focus on the most

relevant features provided by each expert for constructing the final output representation.

Next, the weighted input features W_{att} are combined with the general encoder output G_h :

$$S_h = W_{att} + G_h \quad (7)$$

To perform emotion analysis, a probability distribution over seven emotion classes is generated, indicating the likelihood of the input context C belonging to each emotion class E_{att} . This distribution quantifies the degree of association between the input context and each emotion class.

$$E_{att} = \text{Softmax}(S_H) \quad (8)$$

To improve the performance of the sentiment-aware encoders, the cross-entropy loss is calculated between the predicted emotion label E_{att} and the actual emotion label y_e .

$$L_2 = -y_e \log(E_{att}) \quad (9)$$

3.2.2. Stage 2: Generating Empathetic Response

In this phase, we adopt the techniques of a mixture of empathetic listeners [Lin et al. \(2019\)](#) to generate an empathetic response. The objective is to train each decoder, represented as D_i , to respond to the input context with a specific emotion. These decoders consist of standard decoder layer blocks and take the response embedding Emb^R as the query, and the output of the sentiment-aware encoders S as key and value parameters. Each decoder is assigned a weight based on the probability distributions E_{att} . The outputs of all decoders are then combined and fed into the meta-decoder D_M , which aggregates multiple perspectives from emotion-based decoders to generate a response that appropriately expresses the desired emotion.

$$D_i = \text{Dec}(Emb^R, S_h) \quad (10)$$

$$D_M = \text{Dec}\left(\sum_{i=1}^n D_i E_{att}\right) \quad (11)$$

The negative log likelihood loss is computed by considering the loss between the predicted words with the true words.

$$L_3 = -\log(p(W|R)) \quad (12)$$

Since all components of the proposed model are trained end-to-end, the gradient is computed based on the total of the three loss functions to minimise the sum of individual task losses.

$$L = L_1 + L_2 + L_3 \quad (13)$$

4. Experimental Setup

4.1. Datasets

The experimental setup involves training the proposed model on two datasets: the English Empathetic Dialogue (ED) dataset and the DailyDialog (DD) dataset [Rashkin et al. \(2019\)](#); [Li et al. \(2017\)](#). ED and DD consist of approximately 25,000 and 12,000 multi-turn conversations, respectively. The ED dataset provides annotations for 32 emotions, which we map to the eight basic emotions and two complex emotions defined by Plutchik's wheel of emotions [Plutchik \(1980\)](#). Appendix B provides additional statistics on the datasets.

DD dataset, on the other hand, is already labeled with the six main emotions defined by Plutchik's theory: anger, disgust, fear, happiness, sadness, surprise, and no emotion. By using Plutchik's Wheel of Emotions, we adopt a more manageable framework for modeling emotions compared to considering the 32 emotions individually. This allows for a standardized set of emotion categories across datasets and applications, facilitating the integration of both ED and DD datasets. Complex emotions, which are combinations of basic emotions, are also considered. For example, joy and trust combine to form love, while sadness and disgust combine to form remorse.

By incorporating complex emotions, the model can capture a broader range of emotions, enhancing its sensitivity and the chatbot's ability to provide empathetic and supportive responses. The chatbot can learn to identify emotions that are actually a mix of two or more. In addition, emotions intensify as they move from the outside to the center of the wheel of emotions. For instance, anger ranges from annoyance to rage. As a result, we aggregate all the levels into its basic class. For example, emotions such as "grateful," "excited," and "content," are mapped into the class "joy."

Using Plutchik's emotions provides several advantages. It assists in standardizing emotion categories across datasets. Incorporating complex emotions such as love and remorse into emotion classification may present challenges in prediction accuracy since these emotions may be wrongly labeled as other emotions such as joy or sadness. However, incorporating such emotions broadens the spectrum of emotions captured and improves the chatbot model's sensitivity. This enables chatbots to provide more empathetic and supportive responses, increasing user satisfaction and engagement.

To handle the multi-dimensionality and contextual factors influencing emotions, emotion analysis is performed on top of sentiment analysis. In the ED dataset, emotion classes are classified as positive or negative. Since the DD dataset contains

the "no emotion" class, the classes have been divided into positive, negative, and neutral sentiment classes¹. This approach helps disambiguate emotionally charged emotions, such as love, which may express both positive and negative emotions. By incorporating sentiment analysis into emotion classification, the overall accuracy of emotion analysis is improved.

4.2. Implementation Details

We utilized a virtual instance hosted on the Google Cloud Platform. This instance was equipped with 40 GB of RAM and a single NVIDIA A100 GPU unit. The number of trainable parameters in the model is 74 million. Our model's training took approximately 20 hours to complete. Our approach involved leveraging preexisting code as a basis and building upon it to develop our solution using the PyTorch framework [Lin et al. \(2019\)](#). We used a batch size of 16 and implemented the early stopping strategy. In addition, batch matrix multiplication is employed to improve the model's efficiency. To initialize the embedding, we used transfer learning techniques by employing 300-dimensional English-pretrained Glove embeddings [Pennington et al. \(2014\)](#). The model was optimized with the AdamW optimizer and trained for about 20K steps. Each encoder and decoder consists of 12 layers, with 10 attention heads per layer. In addition, we assigned a weight of 2.5 to the current turn to prioritize its importance. The weight matrices are initialized using Xavier initialization. For more implementation details of our ASEM model, please refer to Appendix A.

4.3. Baselines

Three recent models are used as baselines for comparison with the ASEM model: **(1) CASE** [Zhou et al. \(2023\)](#) model aligns cognition and affection in empathetic dialogue generation by leveraging a commonsense cognition graph and an emotional concept graph. It integrates three stages: graph encoding, coarse-to-fine alignment, and an empathy-aware decoding. **(2) Mixture of empathetic listeners (MoEL)** [Lin et al. \(2019\)](#) model is trained on the ED dataset, which includes 32 emotion classes, with the goal of generating empathetic responses [Rashkin et al. \(2018\)](#). It uses a mixture of decoders, where the model attends to a specific decoder to generate an appropriate empathetic response. **(3) Multitask Transformer (MultiTRS)** [Rashkin et al. \(2018\)](#) model is based on a transformer architecture and incorporates an additional component for emotion classification. This enables the model to classify encoded information into specific emotions

and generate a response based on the encoded information.

4.4. Evaluation Metrics

For the evaluation of chatbots, we followed the framework introduced in [Sedoc et al. \(2019\)](#) for conducting automatic evaluation, and the framework introduced in [Sabour et al. \(2022\)](#) for conducting human evaluation. We evaluate the performance of our model using the following metrics: **BLEU score** [Papineni et al. \(2002\)](#): It measures the similarity between the generated responses and the reference responses using n-gram precision. **Lexical diversity (distinct-n)**: It calculates the average number of unique n-grams in the generated responses, which indicates the diversity of vocabulary used. **Average cosine similarity**: It measures the semantic similarity between the embeddings of the generated responses and the reference responses [Liu et al. \(2016\)](#). **Perplexity**: It quantifies the probability of the model predicting the correct response [Zhang et al. \(2018\)](#).

In addition, we use the macro-F1 score metric to evaluate the emotion classification. Due to imbalanced datasets, we exclude the "no emotion" class when evaluating the performance of the proposed model fine-tuned on the DD dataset. This exclusion is necessary because the dataset contains a large number of dialogues without emotion compared to other classes. While BLEU score is commonly used in chatbot evaluation, it may not be reliable for open-domain response generation since the gold response is not necessarily the only correct response. Therefore, we report the BLEU score as a reference [Liu et al. \(2016\)](#).

In our experiments, we adopted a sequential analysis, first focusing on sentiment analysis before transitioning to emotion analysis. This strategic progression allowed us to harness sentiment data to dynamically refine attention parameters for detecting emotions, thus providing a nuanced insight into their interconnected nature. On the other hand, the joint learning framework implies simultaneous learning where both sentiment and emotion adjustments to attention parameters are conducted in an integrated manner, without the explicit, step-by-step dependency featured in the sequential approach. While joint learning presents benefits in concurrent sentiment-emotion recognition, our sequential methodology stands out, emphasizing sentiment-driven adjustments in attention for enhanced emotion discernment, ultimately yielding superior outcomes.

For human evaluation, we conducted an aspect-based pairwise preference test where we randomly sampled 135 dialogues from the ED test set and 100 from the DD test set. For a given context, we paired our model's response with a response from

¹Preprocessed data is available at

Model	PPL	BLEU	D-1	D-2	Cos.	F1
ED dataset						
MoEL	54.9	2.8	38.6	6.7	24.7	54.4
MultiTRS	43.4	2.4	30.0	4.7	23.5	52.1
CASE	36.7	2.0	29.1	2.0	23.5	52.3
ASEM(j)	42.7	1.63	38.7	2.6	22.4	46.4
ASEM(s)	48.2	2.1	40.0	6.7	24.3	60.6
DD dataset						
MoEL	47.4	1.6	27.1	11.2	18.3	34.2
MultiTRS	43.8	1.7	36.9	9.1	17.5	16.3
CASE	57.8	0.51	42.1	0.68	13.4	30.0
ASEM(s)	43.4	2.2	48.8	9.8	18.4	39.2

Table 2: Automatic Evaluation results. Here PPL denotes perplexity, D-1 and D-2 denote Distinct-1 and -2, Cos. denotes the cosine similarity. 'ASEM (s)' refers to sequential training and 'ASEM (j)' to joint training.

the baselines and presented them to three annotators. The annotators were asked to choose the better response based on the context and three aspects: coherence (Coh.) to assess which response was more coherent in content and relevant to the context, empathy (Emp.) to assess which response showed more understanding of the user's situation and presented a more appropriate emotion, and fluency (Flu.) to evaluate the response's flow and grammar. When both responses are equally good, then the annotator will select "it's a tie". When calculating wins, losses, and ties, we considered the judgements of all three annotators for each of the samples. That is, for each sample, we have three separate evaluations. This method ensures comprehensive evaluations by considering the independent judgements annotators on every sample, as our primary aim was not to achieve consensus but to understand the diverse perspectives of individual annotators.

5. Results and Analysis

5.1. Automatic Evaluation

As shown in Table 2, the ASEM model outperforms baselines in three metrics using the ED and five metrics using the DD datasets. The results suggest that the ASEM model is able to produce a larger variety of unigrams compared to the baselines. However, when considering bigrams, our model generates comparable unique bigrams using the ED dataset, but it does not generate as diverse or unique combinations as the other baselines using the DD dataset, which could result in a higher repetition of word pairs. In addition, the improvement in PPL using the DD dataset indicates that the ASEM model is more confident in predicting a given sequence of words and is able to generate more coherent and fluent responses that are more closely

Comparisons	Aspect	Win	Lose	Tie	κ
ED dataset					
ASEM vs. MoEL	Coh.	45.4	21.2	31.1	0.41
	Emp.	37.3	22.5	40.0	0.44
	Flu.	11.4	3.7	84.7	0.45
ASEM vs. MultiTRS	Coh.	51.1	18.0	30.9	0.47
	Emp.	48.1	17.5	34.3	0.45
	Flu.	12.1	5.4	82.5	0.46
ASEM vs. CASE	Coh.	52.0	15.0	41.0	0.55
	Emp.	71.0	39.0	25.0	0.57
	Flu.	42.0	27.0	66.0	0.65
DD dataset					
ASEM vs. MoEL	Coh.	53.7	33.3	13.0	0.39
	Emp.	37.3	24.7	38.0	0.58
	Flu.	35.0	25.3	39.7	0.42
ASEM vs. MultiTRS	Coh.	46.7	29	24.3	0.33
	Emp.	42.3	18.7	39.0	0.46
	Flu.	17.7	16.0	66.3	0.52
ASEM vs. CASE	Coh.	30.0	35.0	35.0	0.37
	Emp.	29.0	28.0	43.0	0.70
	Flu.	2.0	14.0	84.0	0.88

Table 3: Human evaluation results(%) on the three aspects.

aligned with human-like language usage. Regarding emotion classification using the macro F1 score, the ASEM model significantly outperforms the baselines by improving performance by 6.2% and 5% using the ED and DD datasets, respectively. This improvement suggests that incorporating emotion analysis on top of sentiment analysis unveils the confusion between various emotion classes that can be conveyed through both negative and positive sentiments. Additionally, the findings imply that the model learns an effective representation of the input sequence by employing multiple expert encoders and adding the attention score. By incorporating this score, the model enhances emotion prediction. ASEM improves performance in eight of ten emotion classes significantly. It also removes the confusion between more emotion classes of different sentiments, such as "surprise," which was never predicted as "remorse," and "disgust," which was never predicted as "joy," while the MoEL struggles with emotion classes of different sentiments. This indicates ASEM's capacity to minimize both false positives and false negatives and capture patterns and contextual cues that, in turn, help improve feature representations and generate more empathetic responses (Appendix C shows confusion matrices for the models).

5.2. Human Evaluation

As shown in Table 3, the human A/B test confirmed that our model's responses are preferred by human judges, indicating that the ASEM model outperforms the baselines in all three aspects. Fleiss' kappa measures inter-annotator agreement, with values of $\kappa > 0.4$ and $\kappa > 0.2$ indicating moderate and fair agreement, respectively.

Context	S1: Two years ago i got diagnosed with ms.
	S2: What 'd ms ?
	S1: Its a disease that attacks the nervous system. It was really scary to find out i had it. I just went blind in my right eye one day. Two days and several test later i found out i had ms.
Speaker emotion:	Surprise
MoEL	I am so proud of you ! I hope you had a good experience.
MultiTrs	I hope it turns out well for you
ASEM	I 'm so sorry to hear that. I hope it all works out for the best for you.
Golden	Well , I pray you get through it
<hr/>	
Context	S1: When i lived in california my farmer friend and landlord had me get organized and have all my valuables together in case of a big earthquake. He stockpiled such things as batteriessand water.
	S2: I can understand that as i live in Florida and we prepare for hurricanes each year with a plan and emergency rations and supplies.
	S1: It feels good to get everything together , doesnot it ? you feel as if you are ready for anything !
Speaker emotion:	Anticipation
MoEL	Wow , that is a lot of deal . It is good that you have such a thing in your life
MultiTrs	That is good to hear. I am glad he 's able to help you out. Are the best in life
ASEM	That 's good that you were prepared. Can be so scary !
Golden	Yes, you are correct. You can always hope for the best, But also plan for the worst. Being prepared is a comfortable feeling

Table 4: Two cases of ASEM and the baselines' responses. Blue words indicate ASEM's expected labels.

Context	I never expected him to gift me that beautiful necklace for our anniversary!
Responses	
Sent.Only	It's nice you have someone to look forward to it
Emo. Only	Well, I am glad you have a great time! I am sure you can enjoy it.
Sent+Emo	That is amazing, you must have a very happy life!

Table 5: Results for the Detection Effect of Sentiment and Emotion.

5.3. Case Study

Table 4 displays the responses of ASEM and other baseline models in relation to two distinct emotional states. In the first scenario, MoEL failed to grasp the intended meaning of the user's statement, resulting in an irrelevant empathetic response with an incorrect emotion. Conversely, MultiTRS partially recognized that something was amiss while still generating a neutral and unhelpful response. ASEM effectively apprehended the gravity of the situation, acknowledging the user's emotions with an apology and offering emotional support with positive sentiment by responding with "I hope". In the second case, the user expressed anticipation towards a forthcoming event, and ASEM responded with "were prepared," validating that the model accurately captured the intended emotion and demonstrated a precise understanding of the meaning by replying, "Can be so scary". On the contrary, MoEL once again misunderstood the underlying significance of the last turn, and while MultiTRS acknowledged that the users received assistance from a farmer, it still deviated from the intended meaning, leading both baseline models to perceive

it as something positive. Therefore, ASEM outperforms the baselines in terms of demonstrating empathy and understanding.

We have conducted additional experiments to illustrate the effects of using sentiment detection alone (Sent.), emotion detection alone (Emo.), and the combined use of emotion on top of sentiment detection (Sent+Emo). The results obtained from these experiments provide substantial support for the claims made in the paper as shown in Table 5

5.4. Ablation Analysis

To verify the effectiveness of each model component, we conducted ablation studies, as shown in Table 6:

- **W/O weighted concatenation (We. Conc.):** We observe a decline in context-appropriateness, lexical diversity, emotion classification, and Dist-2 score when disabling the weighted concatenation (We. Conc.) that gives more weight to the current turn. This indicates the importance of considering the current turn in generating appropriate and diverse responses.
- **W One Enc-Dec:** Employing a mixture of experts with multiple encoders and decoders enhances the chatbot's ability to capture sentiment, generate emotionally aware responses, provide response variety, and improve context understanding. Ablating this component and employing only one encoder and one decoder leads to a significant drop in performance across these aspects.
- **W/O sentiment loss (S. Loss):** The absence of sentiment loss results in a decrease in

Model	PPL	BLEU	D-1	D-2	Cos.	F1
ASEM	48.2	2.2	40.0	6.7	24.3	60.6
w/o We. Conc.	48.3	2.4	21.9	5.4	23.5	54.0
One Enc-Dec	48.1	2.1	20.0	5.0	23.7	53.0
w/o S. Loss	57.2	2.1	40.0	6.5	23.0	53.3
w/o S. Enc.	49.4	1.6	42.4	5.0	23.0	48.5

Table 6: Ablation study of our proposed model ASEM using the ED dataset. The best results are highlighted in bold.

macro F1 score, indicating the importance of incorporating sentiment loss in generating empathetic responses. The increase in perplexity suggests a decrease in coherence and difficulty in producing meaningful and relevant responses.

- **W/O sentiment aware encoders (S. Enc):** Removing the sentiment-aware encoders impacts emotion classification, Dist-2 score, and PPL. Sentiment-aware encoders help the model developing effective feature representations and understanding diverse sentiment categories. The absence of this component hinders the model’s ability to extend its understanding and generate appropriate responses.

Overall, the ablation studies confirm the effectiveness of each component in the ASEM model and highlight their contributions to improving various aspects of the model’s performance.

6. Conclusion

In this paper, we propose the ASEM model, which incorporates emotion analysis into sentiment analysis to improve the performance of an empathetic chatbot. We utilized the attention mechanism to focus on relevant aspects of the context while learning the feature representation. Experimental results on two empathetic datasets demonstrate that our model outperforms the baselines. Ablation analysis further confirms the importance of each component in our model. In future work, we plan to explore other attention mechanisms to investigate their impact on generating effective embeddings and improving chatbot performance.

Limitations

Our work has some limitations. Firstly, the lack of a specific-task automatic metric makes it challenging to objectively measure and compare different chatbot models. Secondly, our model has a large number of parameters (around 74 million), which requires a significant amount of training data to achieve optimal performance. However, the available datasets used in our study may not be exten-

sive enough, resulting in repetitive responses and limited diversity or creativity.

Ethics Statement

The ED dataset [Rashkin et al. \(2019\)](#) is sourced from an open-source dataset and involves conversations focused on specified emotions conducted by crowd-sourced workers. The DD dataset [Li et al. \(2017\)](#) was obtained from online platforms targeting English language learners. These conversations do not involve personal privacy concerns. We ensure anonymity in the human evaluation process to maintain ethical standards.

Acknowledgments

This work was made possible by NPRP13S-0112-200037 grant from Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

7. Bibliographical References

- Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. [Multi-task learning for multi-modal emotion recognition and sentiment analysis](#). 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)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.
- Simon Baron-Cohen. 2006. Empathy. *Psychologist*, 19(9):536–537.
- Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. [Specializing static and contextual embeddings in the medical domain using knowledge graphs: Let’s keep it simple](#). In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 69–80, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- 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*, pages 353–363, Aberdeen, Scotland, UK. Association for Computational Linguistics.

- Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. 2019. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. 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)*, Hong Kong, China. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Kim Cheng Patrick Low. 2012. Being empathetic, the way of confucius. *Educational Research (ISSN: 2141-5161) Vol, 3(11):818–826*.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. [MIME: MIMicking emotions for empathetic response generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.
- Tarek Naous, Christian Hokayem, and Hazem Hajj. 2020. [Empathy-driven Arabic conversational chatbot](#). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 58–68, Barcelona, Spain (Online). Association for Computational Linguistics.
- Marret K Noordewier and Seger M Breugelmans. 2013. On the valence of surprise. *Cognition & emotion*, 27(7):1326–1334.
- 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*, pages 311–318.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy.
- 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 Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.
- João Sedoc, Daphne Ippolito, Arun Kirubakaran, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. [ChatEval: A tool for chatbot evaluation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yujia Song. 2015. How to be a proponent of empathy. *Ethical Theory and Moral Practice*, 18(3):437–451.
- Hao Wu, Shaowei Qin, Rencan Nie, Jinde Cao, and Sergey Gorbachev. 2021. Effective collaborative representation learning for multilabel text categorization. *IEEE Transactions on Neural Networks and Learning Systems*.
- Emmanouil Zaranis, Georgios Paraskevopoulos, Athanasios Katsamanis, and Alexandros Potamianos. 2021. Empbot: A t5-based empathetic chatbot focusing on sentiments. *arXiv preprint arXiv:2111.00310*.

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?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. [CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237, Toronto, Canada. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Ling Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307.

A. Implementation Details

For training our ASEM model, we utilized a virtual instance hosted on the Google Cloud Platform. This instance was equipped with 40 GB of RAM and a single NVIDIA A100 GPU unit. The number of trainable parameters in the model is around 57 million. Our model’s training took approximately 20 hours to complete. Our approach involved leveraging preexisting code as a basis and building upon it to develop our solution using the PyTorch framework [Lin et al. \(2019\)](#). To maximize training efficiency, we used a batch size of 16 and implemented the early stopping strategy. In addition, batch matrix multiplication is employed to improve the model’s efficiency. To initialize the embedding, we used transfer learning techniques by employing

300-dimensional English-pretrained Glove embeddings [Pennington et al. \(2014\)](#). The model was optimized with the AdamW optimizer and trained for about 20K steps. Each encoder and decoder consists of 12 layers, with 10 attention heads per layer. In addition, we assigned a weight of 2.5 to the current turn to prioritize its importance. The weight matrices are initialized using Xavier initialization.

In the beam search decoding process employed during inference, the query passed to the decoder at each time step comprises the tokens that have been predicted in the preceding steps. Specifically, for the first step, the query is initiated with a start-of-sequence token, indicative of the beginning of a new response. For each subsequent step, the decoder receives the sequence of tokens that have been generated up to that point. This sequence is used to condition the prediction of the next token, enabling the model to generate a contextually relevant and coherent response progressively.

For emotion analysis, we produced a probability distribution across seven emotion categories. The choice of this specific number is a flexible hyperparameter within our methodology. This parameter allows us to customise the analysis based on the requirements of our study or the specific dataset we are examining. Consequently, while the ED dataset was mapped to 10 emotions and the DD dataset contains 6, a probability distribution over seven emotion classes is generated for the ED dataset and two experts for sentiments, while for the DD dataset, four experts represent different emotions with higher probabilities and three experts for sentiments. We ensured our model did not overfit by using regularization techniques and early stopping during training. These methods prevent the model from learning noise and patterns in the training data that do not generalize well.

B. Dataset

The conversations in the ED were conducted between pairs of human participants. One participant shared a personal narrative, and the other responded empathetically [Rashkin et al. \(2019\)](#). The dataset provides annotations for 32 emotions, which we map to the eight basic emotions and two complex emotions defined by Plutchik’s wheel of emotions, as shown in table 7. The theory identifies various secondary emotions, including aggressiveness and optimism. Nonetheless, in our experimentation, we focused on just two secondary emotions (love and remorse) alongside the primary ones. This approach yielded optimal outcomes, as incorporating additional secondary emotions adversely impacted the accuracy of emotion detection. On the other hand, DailyDialog (DD) is a high-quality multi-turn open-domain English dialogue dataset

Emotion	Sent	Original Emotions
Joy	6568	excited/joyful/grateful/content/confident
Surprise	3321	surprised/impressed
Anticipation	3734	anticipating/hopeful/prepared
Love	3412	sentimental/caring/nostalgic
Trust	3271	proud/trusting/faithful
Anger	5297	angry/annoyed/furious/jealous
Disgust	1270	disgusted
Fear	5979	afraid/terrified/anxious/apprehensive embarrassed
Sadness	5084	sad/lonely/devastated/disappointed
Remorse	2314	guilty/ashamed
10	32	

Table 7: ED Dataset Statistics

Label	Afraid
Speaker	I've been hearing some strange noises around the house at night.
Listener	oh no! That's scary! What do you think it is?
Speaker	I don't know, that's what's making me anxious.
Listener	I'm sorry to hear that. I wish I could help you figure it out

Table 8: Example for one conversation from the ED dataset

Li et al. (2017). On average, there are around 8 speaker turns per dialogue, with around 15 tokens per turn. The DD dataset is already labelled with the six main emotions defined by Plutchik's theory: anger, disgust, fear, happiness, sadness, surprise, and no emotion. Thus, we could not add more secondary emotions. We adhered to their predefined data splits to avoid overfitting. Specifically, the ED dataset is divided into 80% training, 10% validation, and 10% testing splits, encompassing approximately 19,533, 2,770, and 2,547 dialogues, respectively. The DD dataset follows a similar split of 80% for training and 10% each for validation and testing, translating to roughly 11,118, 1,000, and 1,000 dialogues in each set.

Speaker/Listener	Emotion
Speaker: "I was scared stiff of giving my first performance."	Fear
Listener: " Were you? your performance was excellent."	Happiness
Speaker:"Thank you for your kindly words."	Happiness

Table 9: Example for one conversation from the DD dataset

C. More Analysis

Figure 4 shows that ASEM improves performance in eight of ten emotion classes significantly. Additionally, Table 10 presents a comparative analysis of response generation based on sentiment, emotion, and their combination, along with explanations for the Integrated Approach (ASEM).

Context	Today, I think today is a day to learn, to hug, to smile, to make someone happy.
Gold Response	Yes, realistically, I wish it were possible to do that everyday, but general stresses of life get in the way.
Emotion	Anticipation
Responses	
Sent.Only	That is nice to hear. Do you have any plans for that?
Emo. Only	It is good to have something to look forward to. What will you do?
Sent+Emo	You must be so excited! , It is nice to have such a great time!
Explanation	This response effectively combines positive sentiment with the anticipation emotion, acknowledging the excitement and the intent to make the day special.
Context	My daughter is doing such a great job juggling so many things after having a baby that was premature.
Gold Response	that's good. I know that would be difficult.
Emotion	Trust
Responses	
Sent.Only	That is great. How old is she?
Emo. Only	That is great, how old is she ?
Sent+Emo	That's so sweet. You must have a very happy life.
Explanation	This response adds a layer of emotion by suggesting a sense of contentment and happiness in life, which is more engaging but may not fully align with the context of trust and admiration for the daughter's resilience.
Context	I felt sooooo nervous going to my last job interview! Nerves must occur in proportion to how badly you * want * something!
Gold Response	I know what you mean. Did you get the job?
Emotion	Fear
Responses	
Sent.Only	good for you. I am sure you will get one.
Emo. Only	Oh, i am sorry. I hope it goes well!
Sent+Emo	Don't worry, you will be fine!
Explanation	This response directly addresses the user's fear, demonstrating emotion-supporting behavior by providing reassurance.
Context	I have been looking after my wife this week. She has the flu.
Gold Response	Oh poor thing, is she at least getting better?
Emotion	Love
Responses	
Sent.Only	That is great! I hope she gets it.
Emo. Only	I am sorry to hear that. I hope she gets better.
Sent+Emo	That's so nice to hear. I hope she feels better .
Explanation	This response, while not asking questions, expresses a positive sentiment and well wishes, aligning somewhat with expressing care.
Context	I was in total shock the night my fiancé proposed to me at a concert.
Gold Response	That sounds like a great place to propose! How did it go?
Emotion	Surprise
Responses	
Sent.Only	Wow, how exciting! Where did you meet?
Emo. Only	That is awesome! Was it your birthday?
Sent+Emo	Wow , that is amazing news . You must have been so excited!
Explanation	This response acknowledges the user's feelings of surprise and excitement.
Context	Hello, I am very upset about the abuse of some people with their pets...Itis unacceptable
Gold Response	I cannot imagine why people would do that. Animals are helpless.
Emotion	Anger
Responses	
Sent.Only	I am sorry to hear that.
Emo. Only	I am sorry to hear that. What happened?
Sent+Emo	I hate when that happens. It's so annoying .
Explanation	This attempt to align with anger could be seen as demonstrating emotion-supporting behavior, albeit less effectively.
Context	I had to dogsit for my dad about a week ago. I had to clean up dog poop at least 10 times; it was awful.
Gold Response	Ohh, cleaning dog poop is always been annoying, yet petting the dog is always amusing.
Emotion	Disgust
Responses	
Sent.Only	Oh no, that is awful. How did your dog get it back?
Emo. Only	Oh my gosh, that is so gross! I am so sorry to hear that!
Sent+Emo	Wow, that is pretty gross . What did you do?
Explanation	This response acknowledges the user's feelings and asks a question that encourages further sharing.

Table 10: Comparative Analysis of Response Generation Based on Sentiment, Emotion, and Their Combination with Explanations for the Integrated Approach.

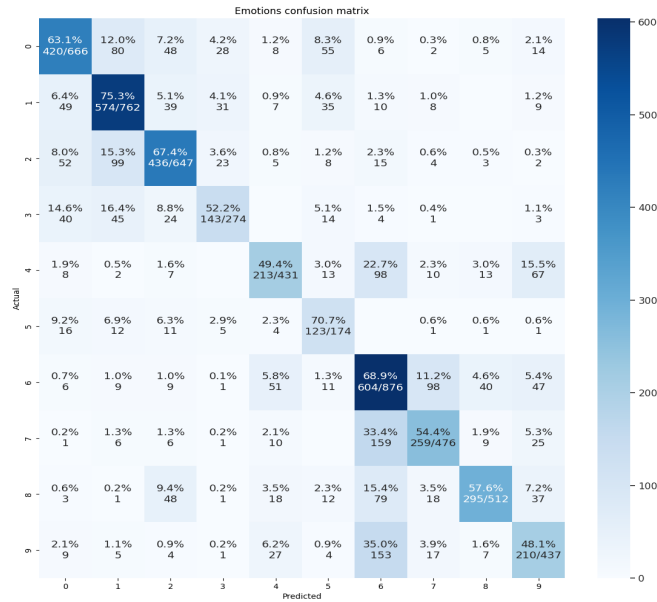
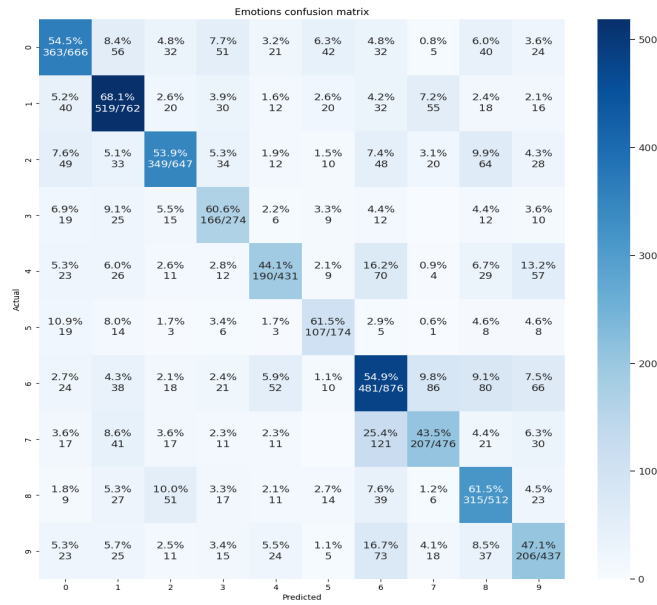


Figure 4: Confusion matrix for MoEL (top) and ASEM (down) models for emotion analysis using ED dataset (0: Anger, 1: Fear, 2: Sadness, 3: Remorse, 4: Surprise, 5: Disgust, 6: Joy, 7: Anticipation, 8: Love, 9: Trust).