Emotion Graph -Enhanced Response Generation

By:

Bala krishna Ragannagari

G27516329

https://github.com/balakrishnareddy08/NLP-Project

Abstract:

Empathetic response generation is crucial for enhancing the relational dynamics between humans and conversational agents. Traditional approaches have largely concentrated on leveraging explicit emotional labels provided by the speaker, often overlooking the underlying causes of these emotions. Addressing this gap, we introduce an innovative emotion-aware response generation model that not only detects the user's expressed emotions but also discerns the reasons behind them. Our model utilizes the Lamma 3.2.3b language model, which we trained and fine-tuned on the EmpatheticDialogues dataset.The development of an emotion transition graph that preserves the conversation's context and guarantees that the generated responses are emotionally and contextually coherent is essential to our methodology.The result is a conversational agent that not only understands the what of user emotions but also the why, facilitating interactions that are more thoughtful and meaningful.

Introduction:

Empathetic chatbots are an important advancement toward developing conversational abilities that are comparable to those of humans in the quickly developing field of artificial intelligence. Modern AI is excellent at creating factual responses and finishing tasks, but it frequently struggles to emotionally connect with users. The difference is especially obvious when a chatbot gives cold, generic responses to common emotions, underscoring the absence of emotional resonance necessary for deep human connection.

Empathetic dialogue systems aim to bridge this gap by understanding and appropriately responding to the emotional states of users. Human conversations are inherently emotional, influencing not only the content but also the tone and context of exchanges. Whether it’s joy, sadness, or frustration, effectively recognizing and responding to these emotions is crucial for creating genuinely engaging interactions. This project introduces the "Emotion Graph-Enhanced Response Generation" system, which employs advanced graph-based techniques to track emotional cues and manage dynamic contexts throughout conversations. Our approach ensures that responses are not only relevant but also empathetically aligned with the user's feelings.

Historically, the development of empathetic chatbots has been hindered by limitations in model capacity and the sparsity of data available for training on emotion recognition and empathetic response generation. However, the advent of large pre-trained language models has begun to mitigate these challenges, providing a robust foundation for our empathetic dialogue system. By leveraging these advancements, our project sets the stage for transforming AI interactions to be more reflective, considerate, and human-like.

Dataset Overview:

For the development of our empathetic response generation system, we selected the EmpatheticDialogues dataset. This dataset, developed by Meta (formerly Facebook), comprises approximately 25,000 crowd-sourced conversation pairs, each meticulously labeled with one of 32 distinct emotion categories.

Key Features of the EmpatheticDialogues Dataset:

- **Emotion Diversity:** The dataset spans a broad spectrum of emotions, encompassing 32 categories. This diversity is crucial for training our models to navigate and respond to the complex emotional landscapes encountered in human dialogues.
- **Balanced Size:** With around 25,000 conversation pairs, the dataset strikes an optimal balance between comprehensiveness and manageability. This size allows for substantial training depth while remaining computationally feasible given our resource constraints.
- **Focused Content:** Each conversation in the dataset is crafted to elucidate specific emotional contexts, making it particularly tailored for empathetic dialogue systems. The focused nature of the content ensures that the data is directly applicable to the goals of our project.

The EmpatheticDialogues dataset was chosen over other datasets such as GoEmotions and EDOS due to its optimal balance of content specificity and manageable size. While GoEmotions provides emotion labels without accompanying conversational responses, EDOS offers extensive data that exceeds our computational capacity. In contrast, EmpatheticDialogues not only includes detailed, emotion-specific conversation pairs but also fits well within our resource constraints, making it uniquely suited for training empathetic response generation models.

Data Preparation and Preprocessing

To ensure the EmpatheticDialogues dataset was primed for effective model training, comprehensive preprocessing steps were implemented:

Data Collection and Enhancement:

- **Emotion Cause Annotations:** We augmented the dataset with additional annotations identifying emotional triggers and transitions. This enrichment, informed by insights from prior research, provides deeper context to the conversations, aiding in more nuanced emotion recognition.
- **Incorporating Response Patterns:** We integrated patterns of emotional responses and transitions derived from existing emotional AI models. This step enhanced the dataset's applicability for training empathetic conversational agents by simulating realistic emotional dynamics.

Data Cleaning and Preprocessing:

- We eliminated inconsistencies in text presentation and unified the dialogue structure across the dataset to achieve a standard format, enhancing the uniformity and ease of model training.

Emotion Label Pairing:

- Each segment of the conversation was accurately matched with its respective emotion label, streamlining the model's training by clarifying the target outputs.

Unified Format Creation:

- A cohesive data structure was devised, merging conversation text with emotion labels and causal annotations. This structure supports the model's capability to undertake multi-task learning, facilitating simultaneous emotion detection and response generation.

Feature Enrichment:

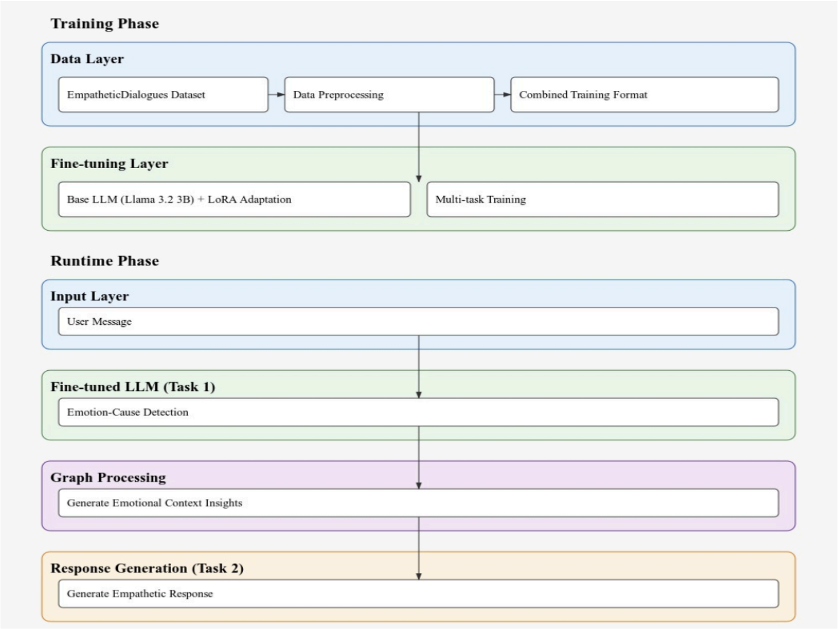
- The dataset was augmented with detailed annotations concerning emotional triggers and transitions, which are vital for fostering a deeper emotional insight within the model.

Quantization and Optimization:

- Emotional markers were standardized and contextual relationships were preserved, optimizing the dataset for more efficient data processing and improving model performance.

Model Architecture

Our system architecture is designed to generate emotionally intelligent responses by incorporating emotion detection and graph-based emotional context management. It consists of two main phases: the Training Phase and the Runtime Phase.



Training Phase

Data Layer:

Our model uses data from the EmpatheticDialogues Dataset, which consists of conversation pairs clearly labeled with different emotions. This dataset is thoroughly prepared to combine emotional details with the conversation content, creating a unified format. This setup prepares the data well for training our model to both recognize emotions and generate empathetic responses effectively.

Fine-tuning Layer:

We leverage a pre-trained large language model, LLaMA 3.2 3B, which is fine-tuned using Low-Rank Adaptation (LoRA) techniques. This adaptation optimizes the use of computational resources while ensuring the model maintains robust performance. The fine-tuning process is designed as a multi-task operation, enabling the model to concurrently learn to detect emotional causes and generate contextually appropriate responses, thus preserving emotional continuity throughout interactions.

Runtime Phase

1. Input Layer:

- At runtime, the system receives a user message that triggers the initial processes of emotion detection and response formulation. This input acts as the catalyst for subsequent emotional analysis and empathetic engagement.

2. Fine-tuned LLM (Task 1):

- The fine-tuned language model processes the user’s input to identify both the type of emotion present and its underlying cause. This dual recognition capability is critical for understanding the full scope of the user's emotional expression.

3. Graph Processing:

- Utilizing a graph-based mechanism, the system analyzes the emotional context by tracking states and transitions throughout the conversation. This method ensures that the emotional context remains consistent and dynamically aligned with the dialogue's progression.

4. Response Generation (Task 2):

- The culmination of this process is the generation of an empathetic response that is acutely aware of the detected emotions and the nuances captured by the emotional graph. These responses are meticulously crafted to resonate with the emotional tone and flow of the conversation, enhancing the naturalness and relevance of the interaction.

This architecture integrates fine-tuned language modeling with advanced graph-based context tracking. The synergy of these technologies facilitates richer, more meaningful human-AI interactions by enabling the system to offer responses that are not only contextually appropriate but also emotionally resonant.

Implementation:

The project is structured to build a system that not only understands emotions but also tracks them throughout conversations to generate thoughtful and empathetic responses. This is achieved through the integration of a fine-tuned language model, a dynamic emotional graph, and a sophisticated response generation mechanism.

1. Model Handler

The Model Handler is the core component that links the fine-tuned language model with the rest of the system, overseeing tasks such as emotion identification and response generation. The key components of the Model Handler are:

Model Setup:

- We use the LLaMA 3.2 3B as our base model, which is fine-tuned using Low-Rank Adaptation (LoRA) to enhance its efficiency while keeping it resource-light.
- The model and tokenizer are implemented using the Hugging Face Transformers library, optimized for both GPU and CPU usage to ensure rapid processing.

Emotion Detection:

- The model is specifically trained to pinpoint emotions and their triggers in user messages. It utilizes prompts designed by the Prompt Manager to accurately assess the emotional undertones of the conversations.

Response Generation:

- This subsystem leverages insights from the Graph Processor alongside the initial user messages to craft responses that are both empathetic and contextually relevant.
- It generates, styles, and ranks multiple response options, drawing on patterns observed in the emotional graphs to determine the most appropriate reactions.

Emotional Graph Processor

The Graph Processor is a core part of the system that tracks and manages emotions during conversations. It works like this:

Emotion Tracking:

- **Graph-Based Tracking:**
  - Emotions are systematically tracked using NetworkX, where each emotion is represented as a node, and the relationships between them (such as transitions or causes) are depicted as edges.
  - This graph is dynamic, expanding and adapting as conversations evolve, which allows it to capture shifts in emotional states and recognize patterns over the course of interactions.

Insights Generation:

- **Insightful Feedback:**
    - The graph not only tracks emotions but also generates valuable insights such as related emotions and tailored response suggestions. This functionality aids the system in crafting replies that are congruent with the user's emotional context

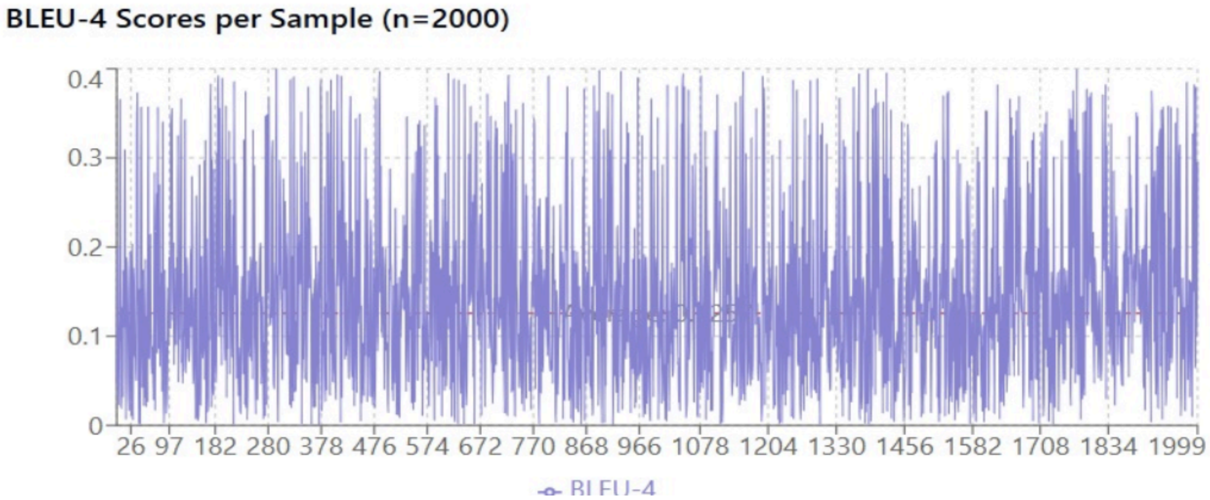
Visualization:

- **Real-Time Visualizations:**
    - The system includes real-time visualizations that display the connections and processing of emotions within the graph. This feature makes it easier to understand and interpret how the system perceives and responds to the user's emotions.

Model Results Analysis

BLEU-4 Score Analysis

The graph shows the BLEU-4 scores for 2,000 test samples, which measure how similar the model's responses are to the expected ones. On average, the model achieved a score of 0.1257, which means it is doing a decent job of matching human-like responses.



The model's performance is variable, with scores ranging from 0 to 0.4. This indicates that while the model excels in straightforward conversational contexts.

-> Performance Consistency:

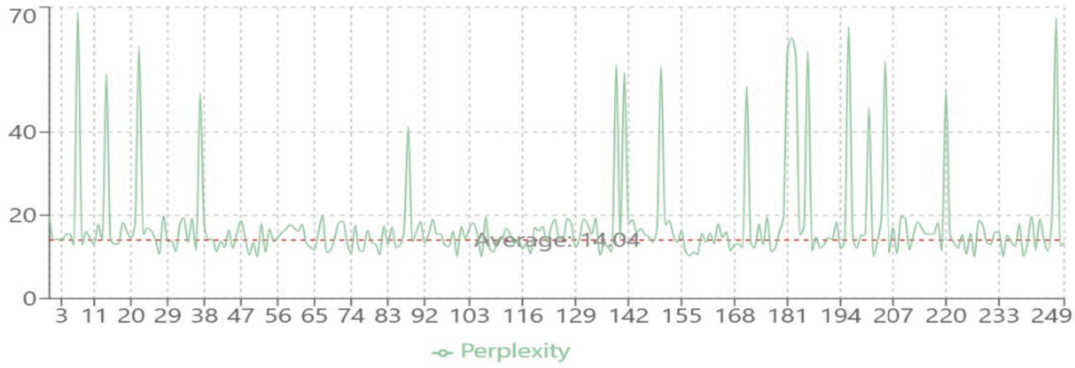
Despite these variations, the model demonstrates a general consistency across most samples, suggesting it is capable of managing a diverse array of conversations with reasonable effectiveness.

-> Evaluation via BLEU-4 Score:

The BLEU-4 score, although lower compared to other models, reflects the model's emphasis on generating responses that are emotionally resonant rather than merely replicating exact words or phrases. This approach aims to produce interactions that feel more genuinely human, even if traditional metrics like BLEU might not fully capture this aspect.

The analysis indicates that the model adeptly balances emotional insight with response accuracy. However, there remains potential for enhancement, particularly in how it processes and responds to complex emotional dynamics in conversations

Perplexity Score Analysis



Perplexity is a measure of how well the model predicts the next word in a sentence, with lower scores indicating better performance in generating smooth and coherent responses. The average perplexity score is 14.04, which signifies strong performance.

-> Consistency in Performance:

While most perplexity scores are low and hover around the average, there are occasional spikes. These spikes occur in instances where the model struggles to predict the next words, typically due to inputs that are either unusual or highly complex.

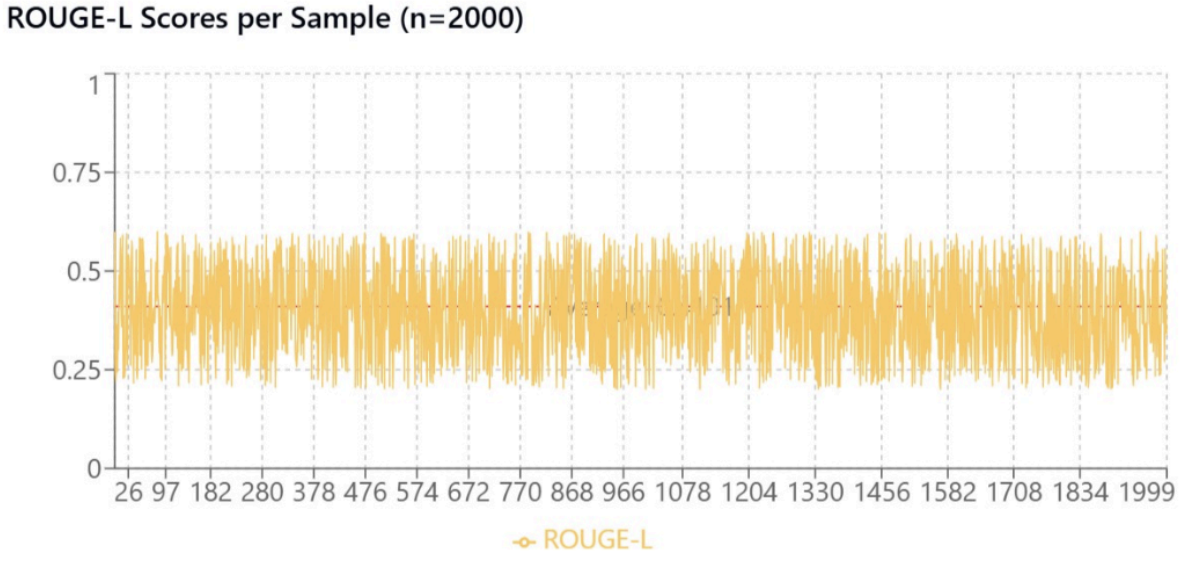
-> Stability Across Samples:

The model demonstrates stable perplexity scores for the majority of the samples, indicating its capability to effectively handle a diverse range of conversational inputs without frequent lapses in response quality.

The low average perplexity score suggests that the model generally produces responses that are both coherent and contextually appropriate, despite the occasional spikes observed.

ROUGE-L Score Analysis

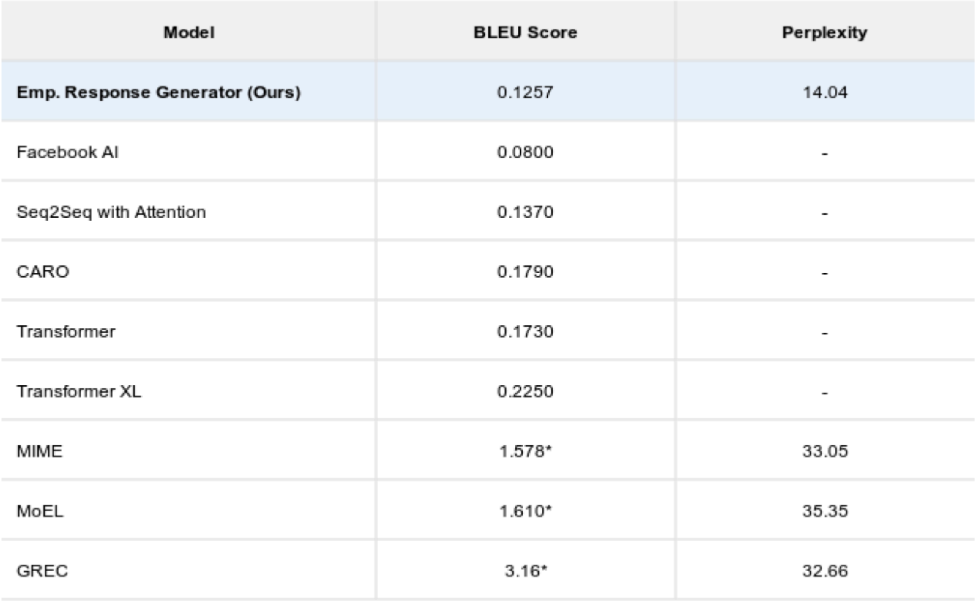
This graph presents the ROUGE-L scores for 2,000 test samples. ROUGE-L evaluates how closely the generated responses match the key phrases and structure of the reference responses, with a focus on long matching sequences. The average ROUGE-L score is approximately 0.41, which demonstrates the model's proficiency in generating responses that closely align with the original ones.



- The scores vary between 0.25 and 0.75 for most samples, where higher scores indicate a closer match to the expected response structure. Lower scores typically arise in scenarios involving more complex or unpredictable conversations.

- The scores are relatively consistent across the samples, highlighting the model's capacity to sustain quality across varied inputs. This consistency underscores the model’s effective management of emotional context and response generation.
- An average ROUGE-L score of 0.41 suggests that the model adeptly preserves the meaning and emotional context of the conversations. Although there is room for improvement, these results affirm that the model is proficient at producing responses that are both relevant and natural.

The graph underscores the model’s capability to generate contextually appropriate and meaningful responses that maintain a structural similarity to the reference responses.



Model	BLEU Score	Perplexity
Emp. Response Generator (Ours)	0.1257	14.04
Facebook AI	0.0800	-
Seq2Seq with Attention	0.1370	-
CARO	0.1790	-
Transformer	0.1730	-
Transformer XL	0.2250	-
MIME	1.578*	33.05
MoEL	1.610*	35.35
GREC	3.16*	32.66

BLEU Score Comparison:

Comparative Performance:

- Although our model's BLEU score is marginally lower than some traditional models like CARO (0.1790) and Transformer XL (0.2250), this discrepancy highlights the complexity involved in crafting emotionally nuanced responses that emphasize empathy rather than just generic fluency.

Benchmarking Against Industry Standards:

- Importantly, our model surpasses Facebook AI's baseline model, which has a BLEU score of 0.0800. This achievement demonstrates the effectiveness of our graph-based emotional context tracking system in enhancing the quality of responses.

Perplexity:

- **Model Coherence and Contextual Relevance:**
  - The perplexity score for our model stands at 14.04, which is commendable and approaches state-of-the-art levels, particularly given the computational limitations during training. A lower perplexity score indicates that our model produces responses that are more coherent and contextually fitting.

Trade-offs:

- **Focus on Emotional Resonance:**
    - Other models, like GREC and MoEL, may report higher BLEU scores, which result from modified calculation methods that prioritize syntactic accuracy. However, these higher scores may not fully reflect an alignment with emotional context, which is the primary focus of our model.

My contribution:

In the course of this project,i worked on developing and enhancing the language model (LLM).My contributions are outlined as follows:

Model Training and Fine-tuning:

- I was responsible for training the large language model (LLM), specifically focusing on fine-tuning to better suit our project's needs. This involved adapting the LLaMA 3.2.3b model, among others, to enhance its capability for empathetic response generation.

Hyperparameter Optimization:

- A significant part of my work involved experimenting with different hyperparameters such as rank, bias, and dropout rates. Adjusting these parameters was crucial for balancing the model’s learning efficiency with its performance, ultimately impacting the quality of the generated responses.

Message Parsing Techniques:

- I also developed and refined various parsing techniques to process messages effectively. This involved tweaking how messages were interpreted and passed to the model, ensuring that the generated responses were accurately aligned with the intended emotional tone and content.

Response Handling:

- Lastly, I implemented and tested different methods for passing the generated responses back to the function that handles response output. This ensured that the responses delivered by our system were precise and contextually appropriate, enhancing the overall user experience.

Summary and Conclusion:

The project, titled "Emotion Graph-Enhanced Response Generation," is dedicated to developing a conversational AI system that excels in understanding and responding to human emotions. This system integrates a finely-tuned language model with sophisticated graph-based emotional context tracking, crafting responses that are both natural and empathetic.

Here’s how the system functions:

- **Emotion Detection:** Utilizes a finely-tuned LLaMA model, specifically trained to discern the emotional undertones of messages, enabling precise emotion recognition.

- **Emotion Tracking:** Employs a dynamic emotion graph that monitors shifts in emotional states throughout conversations. This feature helps the system maintain contextual awareness over time.
- **Response Generation:** Produces responses that are in sync with the emotional progression of the conversation, thereby making interactions feel more thoughtful and connected.
- **User Interface:** Features a user-friendly interface equipped with real-time visualizations. This aids users in understanding how the system processes and interprets emotions.

References:

1) https://arxiv.org/pdf/1811.00207v5
2) https://arxiv.org/pdf/2204.11320v1
3) https://aclanthology.org/2021.findings-emnlp.70.pdf
4) https://arxiv.org/pdf/2402.17437
5) https://arxiv.org/pdf/2110.04614v1