Mental Health Chatbot

Naga Venkata Sahithya Alla
Computer Science and Engineering
University at Buffalo
Buffalo, NY
alla16@buffalo.com

Surya Sumanth Karuturi
Computer Science and Engineering
University at Buffalo
Buffalo, NY
suryasum@buffalo.com

***Abstract*—This project aims to develop an end-to-end conversational system, termed a socialbot, with the capability to engage in open-domain chit-chat and human-like conversations with empathy. The system is designed to receive user input in the form of a query text and generate text responses that are contextually appropriate to the query and ongoing conversation. This project explores the utilization of Large Language Models (LLMs) in the development of a Mental Health Support Chatbot that addresses the mental health crisis by developing an accessible conversational system using Natural Language Processing techniques. The goal is to develop a chatbot that can quickly and meaningfully understand and respond to a variety of emotions and mental health concerns in order to address issues such as people's embarrassment about discussing mental health concerns with others and a lack of resources. This document provides a comprehensive overview of the components and approaches used in the development of our chatbot.**

Keywords—chatbot, mental health, NLP, transformers

I. INTRODUCTION AND PROBLEM STATEMENT

In recent years, there has been significant progress in the development of conversational AI systems known as chatbots. These systems are now widely used across different sectors including education, customer service, personal assistance. Their main purpose is to understand user intentions, maintain coherent conversations, and provide relevant and engaging responses.

Despite these advancements, AI systems still face challenges in effectively handling emotions and providing adequate support for mental health concerns. This project explores the utilization of Large Language Models (LLMs) in the development of a Mental Health Support Chatbot that addresses the mental health crisis by developing an accessible conversational system using Natural Language Processing techniques. Additionally, the project considers the incorporation of Information Retrieval (IR) based response generators to enhance the chatbot's effectiveness in delivering appropriate responses. The aim is to tackle challenges like feelings of shame and resource scarcity by designing a chatbot that can swiftly and empathetically understand and respond to diverse emotional states and mental health issues. The model will also produce facts and chit-chat with the user when required. To develop a conversational system, we will combine an Information Retrieval (IR) model with a neural Language Model (LM) trained on p (R|C, Q). Each time, the system will offer multiple potential responses <R_a, R_b, ..., R_n>, which a neural Dialogue Manager (DM) will review to choose the best one by ranking them. Importantly, all neural LMs will be trained to minimize the language modeling loss between the generated response R and the ideal response Y. This comprehensive approach ensures the system can utilize both traditional IR methods and advanced neural language modeling to produce contextually appropriate and fluent responses during conversations.

II. SYSTEM ARCHITECTURE

The approach we designed for developing mental health chatbot enables it to engage in chit-chat, empathetic conversations while also providing factual information as needed.

The users can interact with the chatbot via a user interface, initiating the conversation with their query. Initially the user's utterance will go through a text cleaning step to remove unnecessary punctuations and convert the text to lower case, following which we have utilized Named Entity Recognition (NER) to identify and record relevant entities. By retaining context, this will be helpful for us to store the conversation history as well. The next step is intent classification, which detects the intent of the user's query. This module will decide if they are looking for an empathetic conversation, chit-chat, or factual information. The query is then passed to the dialogue manager in which we have a context tracker module that keeps track of the entities found using NER and maintains the conversation's flow. The query is then routed to the most appropriate response generator – factual, empathetic, or chit-chat based on the intent that has been identified. We are training the chit-chat generator using T5/ DialoGPT model using BYU PCCL chitchat dataset. To train the Empathetic dialogue generator, we fine-tuned the T5 model on Empathetic Dialogues dataset. We further fine-tuned this T5 model again with Chatbot Mental Health conversations data available on Kaggle for diversity. The factual data generator is trained on the General-Knowledge dataset from hugging face using information retrieval technique which will be further discussed in section E. If the chatbot is unable to identity an intent that it can handle, it will fallback to a message stating that there is not enough information at this time. The generated responses are passed through a neural re-ranker which will assess the generated responses to determine which option is most appropriate. Then, the chosen response is passed through a sentence completion or language polishing module which will further refine the generated response to guarantee that they are coherent, fluent, and grammatically correct. Finally, the response will be displayed on the user interface after these steps.

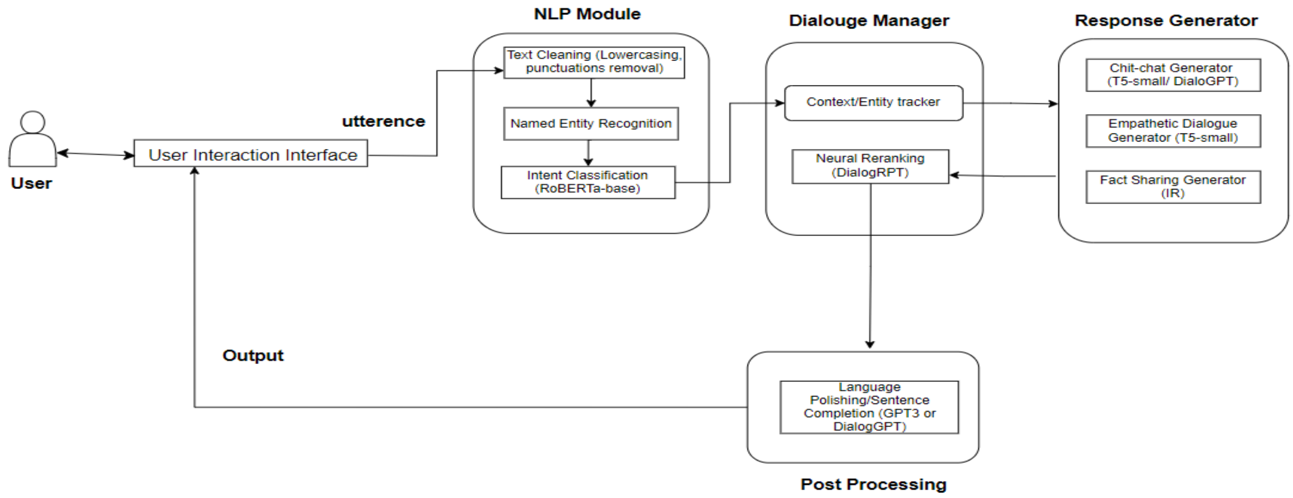

Figure 1: System Architecture

A. *Named Entity Recognition: (In-progress)*

The Named Entity Recognition (NER) module is incorporated to capture entities from the user input utterances and chatbot response, enhancing both context tracking and conversation flow. We are planning to train the NER module on BERT (Bidirectional Encoder Representation from Transformers)/RoBERTa (Robustly Optimized Bert Approach) model. We will be using the CoNLL-2003 dataset. This rich dataset contains news articles from the Reuters Corpus annotated with named entities. We will use this dataset to fine-tune the pre-trained BERT or RoBERTa model which will help us in understanding and maintaining the context.

B. *Intent Classification:*

The objective of intent classification is to identify the underlying intent in a user's utterance, classifying it into one of the three categories: chit-chat, empathetic, or factual. For this task, we employed the RoBERTa-base model for training our classifier, utilizing a composite dataset comprising the Empathetic dialogue dataset, Chatbot Mental Health conversations, BYU PCCL chitchat dataset, and the General-Knowledge factual dataset from hugging face. The RoBERTa-base model works by breaking down the text/utterance into smaller sections, examining each for potential intent, and then combining these evaluations to choose the intent that appears most frequently across the segments.

C. *Empathetic Dialogue Generator:*

The Empathetic Dialogue Generator is trained on the empathetic dialogues dataset which has around 25K conversations by fine tuning the T5 (Text-to-Text Transfer Transformer) model. The model was further fine-tuned using the Chatbot Mental Health Conversations dataset which consists of 37K records, to further improve the accuracy and diversity of response generation.

D. *Chit-chat dialogue Generator: (In-progress)*

This Response generator particularly focus on engaging in casual conversations. The chitchat response generator is a T5/DialoGPT model that has been fine-tuned using the BYU PCCL Chitchat Dataset. We are still running the data on both models individually to compare the results. BYU PCCL dataset is a collection of conversation data by Brigham Young University's Pronoun Correcting Classifier Lab. The dataset contains various conversational data across different topics and contexts. We must preprocess this dataset by combining the previous utterances also to be part of the input and the current utterance as the output. We will use this dataset for the training the model.

We will evaluate the model's performance on the BYU PPCL Chitchat Dataset using the human evaluation. There are also automated metrics for evaluation such as BLUE score, METEOR score, and ROUGE score. These metrics evaluates the performance by measuring the similarity between generated responses and the reference responses. By training on this dataset, it will enhance the model's response generation capability in general conversations.

E. *Factual Information Generator:*

The model leverages semantic similarity computations to match user queries with the most relevant information from the dataset, which is a core concept in Information Retrieval (IR) subject. The dataset used is the General-Knowledge factual dataset from hugging face. We implemented a question-answering system using the sentence transformer all-MiniLM-L6-v2 model for computing the semantic similarity to effectively match the user queries with relevant answers. After loading the pre-trained model and tokenizer, we encoded the input texts into embeddings through tokenization, padding and truncation, followed by the embedding generation via the transformer model and mean pooling. The model calculates the similarity between user query and fact embeddings using cosine similarity and also includes a query matching function that searches the database of facts for highest similarity score to identify the most relevant answer.

F. Neural re-ranker: (In-progress)

We plan to implement a DialogRPT-based re-ranking mechanism to enhance the chatbot conversations quality by providing the appropriate and relevant response. The DialogRPT model, which is built by Microsoft, developed on the GPT-2 architecture, is served as a cutting-edge tool for assessing response appropriateness and quality. It operates by comparing the generated responses likelihood scores with the derived intent from the intent classification system. This method allows for a more nuanced evaluation of potential responses by ensuring that the chatbot delivers a high-quality and contextually relevant interactions.

G. Sentence Completion or Language Polishing module: (In-progress)

This module will refine the fluency and appropriateness of the bot's responses by integrating NLP models. We are planning to try the GPT, BERT, T5 models by fine-tuning them on conversational data to ensure they capture the nuances of the human language. This module will enable the generation of contextually relevant and coherent responses prioritizing quality and grammatical correctness.

III. RESULTS AND IMPROVEMENTS

A. Results of Intent Classifier:

For now, we have trained only on two classes – Empathetic generator data and factual information data. As chit-chat generator training is still in progress, that dataset will be included once the chit-chat generator is up and running. The following results are only based on the remaining two classes:

TABLE I. INTENT CLASSIFIER RESULTS

Metric	*Baseline Results*
Accuracy	0.84
Precision	0.84
Recall	0.83
F1	0.83

The intent classifier built with RoBERTa-base model has good initial performance with an accuracy and precision of 0.84, recall of 0.83, and an F1 score of 0.83 in differentiating between factual and empathetic intents. These metrics show that the system is reliable but can be improved for even better results.

To improve, we can investigate and address the class-imbalances, misclassifications and make sure the model is not biased. We can also check the performance once chit-chat data is also added.

B. Results of Empathetic Dialouge Generator:

TABLE II. EMPATHETIC DIALOGUE GENERATOR PERFORMANCE

Metric	*Baseline Results*
AVG BLEU score	0.04725
BLEURT	-0.95
ROUGE (1,2,L)	0.13, 0.012, 0.118
BERT	0.84
PERPLEXITY	18.21

From the above metrics, we can understand that there is scope for improvement of our model. The Avg BLEU score at 0.047 suggests that the model's responses have limited overlap with the reference responses. But when compared with the scores in the original paper [2] which is 0.062, we are not at a very bad stage, but still has a good scope for improving our model. The BLEURT score of -0.95 highlights difficulties in generating contextually and emotionally appropriate responses showing necessity to enhance the model further. The ROUGE scores for unigrams, bigrams and the longest common subsequence suggests that it is not capturing essential content and phrasing from human references. However, the BERT score of 0.84 implies that the model is capable of producing semantically similar responses even if wording diverges from human responses pointing towards semantic understanding. The perplexity score also indicates room for improvements in model fluency and confidence.

To improve the model, we are planning to run the model with a greater number of epochs and trying out different hyperparameters. Also, improving the quality of response which is generated by using sentence completion techniques could also help us to improve scores and fluency. We can also try to add more data (if we find applicable data and if required) to see if there is any improvement in performance.

Generated Responses:

```
Query I am feeling very sad today'
Generated text: i'm sorry to hear that. i'm here for you. so please talk to me

Query do you how happy i am?'
Generated text: i'm glad you have a great day.

Query I had a big fight with my bestfriend'
Generated text: i'm sorry to hear that. i'm here for you. so how are you feeling today?

Query bye
Generated text: bye! come back again

Query what is anxiety
Generated text: anxiety is a mental health condition that affects how you feel. anxiety can be overwhelming.
```

C. Factual Information Generator:

The response retrieval in factual information generator is quite good as of now as we are comparing the semantic similarity of the sentences. We are also trying out how BM25 model will work on data retrieval process and see if it is generating better results. This can be further improved by incorporating a text summary model to summarize the resulted response as few responses are so long and we want to keep it short and concise for engaging the user. Further, it can be improved in terms of fluency and coherence through language polishing techniques. We have done human evaluation and assessment on this generator as we needed to check and verify if the facts generated by the chatbot are factually correct or not and we have discovered that it is generating pretty good results and has an effective fallback mechanism when it does not have the information.

Generated Responses:

```
Query: i want to know about largest land animal
Response: the african elephant is the largest land animal.
```

```
Query: can you tell me what is the largest species of squid
Response: the colossal squid is the largest species of squid, with specimens that can reach lengths of up to 46 feet and weigh over 1,000 pounds.
```

Fallback Response:

```
Query: do you know how to do integration in mathematics
Response: I'm sorry, I don't have information on that topic.
```

Overall, once the modules that are in progress are built and are integrated, we are aiming to evaluate various models and addition of datasets (if required), tune hyper-parameters and see which models are feasible for our resource environment and employ or continue with those ensuring both efficiency and effectiveness.

Our bot conversation flow:

```
Bot: Hi! How can I help you today?
User: hi
Bot: i'm listening. please go on.
User: i have depression
Bot: i am sorry to hear that. if you are feeling depressed, it is helpful to find out what you're feeling this way.
User: can you help me by giving a suggestion
Bot: i can suggest you to give yourself a break.
User: can you prevent mental health problems?
Bot: it helps to talk about what's happening. you're going to be okay
User: where can i go find therapy?
Bot: therapy is a form of treatment that aims to help resolve mental or emotional issues. it is helpful for those with mental health conditions or even everyday life challenges.
User: thanks
Bot: happy to help!
User: bye
Bot: bye! come back again
User: quit
```

Resources: We are using V100 High-RAM NVIDIA Tesla GPU in Google Colab pro to run our notebooks.

REFERENCES

[1] https://arxiv.org/pdf/2109.02513.pdf
[2] https://arxiv.org/pdf/1811.00207.pdf
[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4] https://github.com/facebookresearch/EmpatheticDialogues/tree/main
[5] https://www.kaggle.com/code/jocelyndumlao/chatbot-for-mental-health-conversations
[6] https://huggingface.co/datasets/MuskumPillerum/General-Knowledge