Large Language Models for Societal Good: Mental Health Chatbot

Naga Venkata Sahithya Alla
Computer Science and Engineering
University at Buffalo
alla16@buffalo.edu

Surya Sumanth Karuturi
Computer Science and Engineering
University at Buffalo
suryasum@buffalo.edu

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.

1 Introduction

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.

2 Related Work

Numerous studies have significantly advanced dialogue generation and empathetic responses in chatbots. Below, we summarize essential works that showcase innovative methodologies and strategies for creating cutting-edge conversational systems: DeepMoji (Felbo et al., 2017) is a deep learning model trained on over 1.2 billion tweets annotated with emojis. By accurately predicting emotional undertones in conversations, it lays a foundational framework for empathetic chatbots to understand and respond with context-sensitive emotional nuances. Microsoft DialogRPT (2021), built on the GPT-2 model, applies neural ranking techniques to evaluate the quality of chatbot responses. By comparing generated responses with the user's conversational intent, it ensures nuanced interactions through meaningful exchanges. Li et al. (2016) employed reinforcement learning to enhance response diversity and reduce repetition in chatbot conversations. Their reinforcement learning framework aimed to maximize long-term conversational success by incorporating user engagement as a reinforcement signal. Rashkin et al. (2019) set benchmarks for empathetic dialogue generation using retrieval and generative models. These models used transformer architectures to balance empathetic and factual responses, relying on emotion and topic detection to guide the generation process across a context window of four past utterances. TransferTransfo (Wolf et al., 2019) utilizes the GPT-2 architecture to apply conversational transfer learning. This method leverages pre-trained models for high-quality response generation and open-domain capabilities. Its transfer learning approach makes conversations more empathetic and contextually relevant. These works collectively emphasize various strategies like large-scale datasets, deep learning, reinforcement learning, and transfer learning. Together, they contribute to continuous improvements in creating more nuanced and empathetic chatbots.

3 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 model using

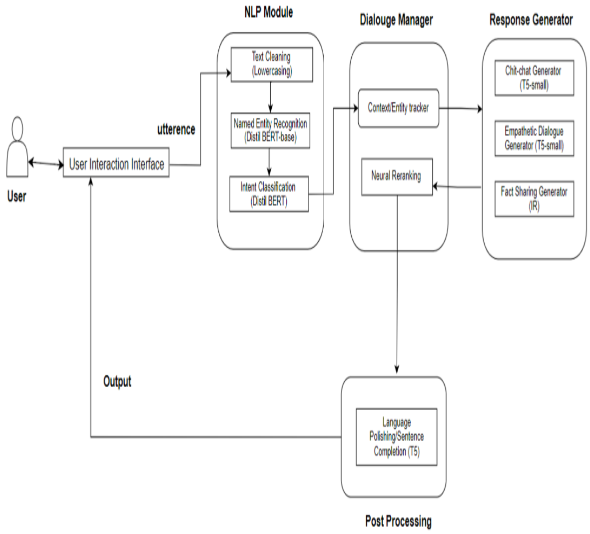

Figure 1: System Architecture

BYU PCCL chitchat the Amazon Alexa Topical Chat 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 3.5. 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.

3.1 Named Entity Recognition

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. This module is developed using the DistilBERT model and fine-tuned with the CoNLL-2003 dataset. This rich dataset contains news articles from the Reuters Corpus annotated with named entities which provided a robust foundation for training our model. This module will help us in understanding and maintaining the context.

3.2 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 DistilBERT-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 DistilBERT-base model processes text by first tokenizing the input utterance into subwords or tokens. It then uses a self-attention mechanism to understand the context around each token and aggregates this information to form a contextual understanding of the entire input. Based on this understanding, it classifies the intent of the utterance into the most probable category among chit-chat, empathetic, and factual. DistilBERT is a version of BERT, which is designed for faster performance with fewer parameters. We haved trained our model for three epochs using a weight decay value of 0.01.

3.3 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 for four epochs with a learning rate of 2e-4 and a weight decay value of 0.01. 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. The model was trained for three epochs with a learning rate of 0.001 and weight decay of 0.001 on Nvidia A100 GPU.

3.4 Chit-chat Dialogue Generator

The chitchat dialogue generator is built to facilitate casual, open-domain, and topical conversations. For this, the T5 model has been carefully fine-tuned using two main datasets: the BYU PCCL Chitchat Dataset and the Amazon Alexa Topical Chat Dataset.The BYU PCCL Chitchat Dataset, curated by Brigham Young University's Pronoun Correcting Classifier Lab, offers conversational data spanning various topics and contexts. Preprocessing of this dataset involved merging continuous conversations from the same user into cohesive conversation records. After preprocessing, the dataset contained 66,000 conversation records, providing a substantial resource for the T5 model to learn from and adapt to diverse casual conversation patterns.The T5 model was further refined using the Amazon Alexa Topical Chat Dataset, which contains 188,000 conversation records. This dataset helps the T5 model engage in topical chats that cover a broad spectrum of dialogue contexts. By leveraging these comprehensive datasets, the T5 model can generate accurate, coherent responses that cater to a wide range of open-domain and topical chat scenarios. To measure the model's performance, human evaluations are complemented with automated metrics like BLEU, BERT, and ROUGE scores. These metrics compare the similarity between generated responses and reference responses, offering quantifiable insights into response generation improvements.In summary, training on these datasets has significantly enhanced the T5 model's ability to handle casual, open-domain, and topical conversations. This ensures that the chitchat dialogue generator provides users with a more natural, relevant, and comprehensive dialogues.

3.5 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.

$$CosineSimilarity(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$$

This model also includes a query matching function that searches the database of facts for highest similarity score to identify the most relevant answer.

3.6 Dialouge Manager

The dialogue manager in our chatbot is essential to keeping users engaged in a logical and contextually appropriate conversation dialogue. In order to maintain a contextual understanding of the continuing conversation, the system stores recent interactions and entities recognized by the NER module. The conversation will be cohesive, and relevant responses are produced because of the stored context. In order to evaluate conversational continuity and make sure that the dialogue flows naturally and logically, the manager also calculates the similarity between successive user utterances. By categorizing each phrase, the dialogue manager uses pre-trained models to determine the user's intents. From a set of pre-defined response generators which are designed for factual, chitchat, and empathetic conversations, it chooses the most appropriate response generator based on the gathered context and the classified intent. In addition, the manager uses a neural reranking module on the responses that are generated. In order to guarantee that the response provided is the most contextually relevant, this reranking module is based on how similar every response is to the user's current query semantically. The conversational experience is improved overall by this systematic procedure, which makes interacting with the chatbot seem more interesting and natural.

3.7 Neural re-ranker

We incorporated a neural reranking mechanism in our system, which is crucial for improving the relevancy of the answers given to the user's ask. The function uses a pre-trained bert-base-nli-mean-tokens model to convert each response into a dense vector i.e., it converts the response's semantic content to a high dimensional space. In a similar way, the user;s utterance is also converted into the matching vector representation. Post this, the cosine similarity between the response vector and query vector are computed where a lower number denotes greater relevance. This distance metric efficiently measures the semantic similarity between the query and each response generated by our main response generator modules. After that, the answers are arranged according to the distances where the response with the most similarity to the user query is ranked higher. The response with the shortest cosine distance, which is the most relevant response, would usually be chosen as the final output. This approach greatly improves the process of selecting responses by making sure that they are both contextually relevant and closely connected with the user's ask.

3.8 Sentence Completion Module

The sentence polishing module improves the fluency and accuracy of the chatbot's responses by integrating Natural Language Processing (NLP) techniques. The primary goal is to ensure that the bot produces coherent and contextually appropriate responses while maintaining grammatical precision and overall quality. The T5 (Text-to-Text Transfer Transformer) model is utilized for training. This model translates various NLP tasks into a unified text-to-text format. Datasets used are PAWS (Paraphrase Adversaries from Word Scrambling) dataset which contains pairs of sentences that are syntactically similar but semantically different. It tests the chatbot's ability to distinguish between these pairs and produce meaningful and contextually appropriate paraphrases or distinguish dissimilar sentences and JFLEG (JHU Fluency-Extended GEC) dataset which is a grammatical error correction dataset aimed at improving response fluency and grammatical accuracy. It contains corrected sentences from non-native English speakers, offering valuable data for enhancing sentence polishing. By fine-tuning the T5 model on PAWS and JFLEG datasets, the bot learns to comprehend linguistic nuances and generate high-quality responses.

4 Results

4.1 Named Entity Recognition

The evaluation of our NER system demonstrated robust performance across several key metrics. The module achieved an accuracy of 0.98 which shows high overall correctness in predictions. The F1 score which is a balanced measure of precision and recall, was recorded at 0.90, illustrating the model's efficiency in balancing both aspects effectively. All metrics are documented above in Table 1.

Table 1: NER Results

Metrics	Final
Accuracy	0.98
Precision	0.89
Recall	0.91
F1 score	0.90

These results underscore the system's suitability for our application as it needs a good precision, and it needs to be reliable as we are saving the entities to track the context.

4.2 Intent Classification

For baseline, we have trained only on two classes – Empathetic generator data and factual information data. As chit-chat generator training was still in progress. The following results obtained for baseline model which was trained using RoBERTA model.

Table 2: Baseline Results using RoBERTa

Metrics	Baseline
Accuracy	0.84
Precision	0.84
Recall	0.83
F1 score	0.83

The final intent classifier was trained on all classes – empathetic dataset, factual information dataset, and chit-chat and topical dataset. This model is built with DistilBERT model. It showed a decent performance for such huge dataset with an accuracy of 0.76, precision of 0.84, recall of 0.76, and F1 score of 0.75.

Table 3: Final Results using DistilBERT

Metrics	Final
Accuracy	0.76
Precision	0.84
Recall	0.76
F1 score	0.75

To improve, we can investigate and address the class- imbalances, misclassifications and make sure the model is not biased.

4.3 Empathetic Response Generator

Table 4: Empathetic Response Generator results

Metrics	Baseline	Final
Avg BLEU score	0.04725	0.05999
ROUGE(1,2,L)	0.13, 0.012, 0.118	0.15, 0.024, 0.134
BERT score	0.84	0.85
Perplexity	18.21	2.66

From the metrics in table 4, we can see that there is improvement when compared to the baseline results with average BLEU score at 0.0599 which suggests that the model's responses have limited overlap with the ground truth responses. But when compared with the scores in the original paper [2] which is 0.062, we are at a good stage comparatively and improved our scores from baseline results. 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 0.85 implies that the model is capable of producing semantically similar responses even if wording diverges from human responses pointing towards semantic understanding. There is a slight improvement in this score as well when compared to baseline. There is a major improvement in the perplexity score being 2.66 for final model which suggests that the model is fluent and confident.

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.

Results:
Input: I am feeling sad
Generated Response: I'm sorry to hear that. what's the reason behind this?

Input: I got promoted
Generated Response: That's great. how did you feel feeling today?

Input: I broke up with my bestfriend. How do i overcome the depression? Can I go out and watch a movie?
Generated Response: That's a great way to deal with the depression. I'm here for you. If you're feeling this way, you can talk to me.

Input: I am very excited today!
Generated Response: What do you think is the reason behind this?

4.4 Chit-Chat Response Generator

The results from the chitchat dialogue generator demonstrated positive progress across key metrics. An average BLEU score of 0.13922 showcases a meaningful similarity between generated and reference responses. ROUGE scores (1: 0.13979, 2:

0.15832, L: 0.11590) indicate that the model effectively captures essential content and structure. The strong BERT score of 0.82501 reflects high semantic understanding, while a perplexity score of 1.16895 suggests confident, fluent response generation.

Table 5: Chit-Chat Response Generator results

Metrics	Final
Avg BLEU score	0.1392
ROUGE(1,2,L)	0.13979, 0.15832, 0.1159
BERT score	0.82
Perplexity	1.168

Results:
Input: I did scuba diving today
Generated Response: Thats awesome i havent done scuba diving yet

Input: How is your evening
Generated Response: Pretty good

Input: Hello
Generated Response: Hello there. tell me how are you feeling today?

4.5 Factual Response 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 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.

Results:
Input: Name 4 sports that are typically played in summer
Generated Response: Swimming, volleyball, softball, and tennis are all sports typically played in the summer.

Input:do you know about any movies?
Generated Response: 1. The Lord of the Rings by J.R.R. Tolkien
2. 1984 by George Orwell
3. Harry Potter by J. K. Rowling
4. Gone With the Wind by Margaret Mitchell
5. To Kill a Mockingbird by Harper Lee

Input: what is summer
Generated Response: To me, summer represents a time of fun and relaxation, full of sunny days and warm nights

Fallback:
Input: what is ABC
Generated Response: Sorry, I don't have information on that topic.

4.6 Sentence Polishing Module

The sentence polishing module indicated promising results across various metrics. The average BLEU score of 0.86394 suggests a high overlap between polished sentences and reference texts, demonstrating strong fluency. The ROUGE scores (1: 0.88341, 2: 0.67141, L: 0.78324) show that the module effectively captures essential phrases and structure. A BERT score of 0.96175 reflects a high degree of semantic similarity, indicating that polished sentences convey meanings similar to the references. Lastly, a perplexity score of 3.17043 indicates confident response generation with low uncertainty.

Table 6: Sentence Polishing Module results

Metrics	Final
Avg BLEU score	0.8639
ROUGE(1,2,L)	0.883, 0.671, 0.783
BERT score	0.96
Perplexity	3.17

5 Discussion and Error Analysis

From the above results, we can see that our chatbot demonstrated strong performance across modules. The NER module achieved high accuracy which is crucial module for context tracking. However, we can improve the model significantly by training the modules for more epochs if computational resources are given. To improve the intent classification results, we need more robust data for factual information as our model is getting confused sometimes while classifying as our factual data is general knowledge based and is closer to the chit-chat conversations by being more casual. We can also improve this by addressing the class imbalances to make sure the model is not biased. The empathetic response generator showed good results but it can be further improved by having more diverse dataset and attribute (feeling) based conversation handling. The chit-chat generator performed well, suggesting effective casual conversation handling. The factual response generator performed well in delivering accurate information, showing its effectiveness in knowledge-based responses. The sentence completion module can be further improved to tailor the responses into more cohesive and fluent response with good punctutions given more dataset. The context management can be further improved by using memory networks or more advanced frameworks for better results and conversation flow.

6 Conclusion

In conclusion, the unique integration of multiple advanced models like BERT, DistilBERT, and T5 models, which enhanced the chatbot's ability to handle diverse tasks such as entity recognition, intent classification, and response generation with modules like empathetic, chit-chat/topical, and factual information generators more effectively rather than having only single-model system. For further enhancements, we can focus on refining the training process expanding the dataset diversity and also integrate user feedback mechanism to optimize the performance and results.

7 Contributions

Table 7: Contribution Table

Name	Task
Sahithya	Empathetic Response Generator, Factual Response generator NER, Intent Classification and both worked together on other modules and report.
Surya	Chit-chat Response Generator, Sentence Polishing Intent Classification and both worked equally on other modules and report.

References

[1] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. https://arxiv.org/abs/1708.00524.

[2] Xiang Gao, Yizhe Zhang, Michel Galley,Chris Brockett, Bill Dolan. 2020 Dialogue Response Ranking Training with Large-Scale Human Feedback Data https://arxiv.org/abs/2009.06978.

[3] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, Dan Jurafsky, 2016 Deep Reinforcement Learning for Dialogue Generation https://arxiv.org/abs/1606.01541.

[4] Hannah Rashkin, Eric Michael Smith, Margaret Li, Y-Lan Boureau. 2019 Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset https://arxiv.org/pdf/1811.00207.pdf.

[5] Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K. Srihari. 2021. Proto: A Neural Cocktail for Generating Appealing Conversations. https://arxiv.org/pdf/2109.02513.pdf.

[6] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019 Transfer-Transfo: A Transfer Learning Approach for Neural Network Based Conversational Agents https://arxiv.org/pdf/2109.02513.pdf.