Next Utterance Prediction For Mental Health Counseling

Wasif Ali(2022583) Ashish Bargoti(2022114) Roshan Kuman Mahto(2022418)

Abstract

This project presents a novel approach to next utterance prediction for mental health counseling dialogues using transformer-based models. The primary objective is to generate empathetic, contextually appropriate responses that can support therapists and enhance scalable mental health care. We propose two distinct models. The first model, *SentimentInfusedBart*, extends the conventional BART encoder-decoder framework by incorporating external sentiment features derived from a pre-trained sentiment analysis module. The second model, referred to as *T-GenSim*, employs a dual-model architecture that combines a generator model for immediate therapist response prediction with a simulator model that forecasts potential follow-up patient inputs. These results underscore the promise of incorporating sentiment-aware and hybrid learning strategies in creating effective dialogue systems for mental health counseling.

1 Introduction

Conversational modeling has made a significant progress with development of large pre-trained language models. Yet, using such models in sensitive fields such as mental health counseling poses distinct difficulties. This activity not only requires comprehension of the context and purpose of earlier statements but also makes it important to generate responses that are empathetic as well as contextually relevant.The aim of this project is to contribute to smart dialogue agents for assisting therapists or creating scalable mental health support tools.

2 Related Work

The integration of large language models (LLMs) into mental health counseling has gained attention for its potential to assist therapists. Inaba et al. (2024) explored GPT-4's ability to generate counseling responses through role-play dialogues. Their study found that GPT-4's responses were comparable to those of human counselors, demonstrating its potential in therapeutic contexts. However, they highlighted the need for improved emotional understanding and personalized responses.

Similarly, Srivastava et al. (2023) introduced READER, a model that uses response-act prediction and reinforcement learning to generate more contextually relevant counseling responses. Evaluated on the HOPE dataset, READER outperformed baseline models in emotional relevance and response appropriateness. This work emphasizes the importance of guiding the response generation process to achieve better contextual awareness.

Both studies underline the promise of LLMs in enhancing therapeutic practices. However, challenges remain in ensuring that these models can generate responses with the necessary emotional sensitivity and long-term coherence for mental health dialogues. Future research should focus on improving the emotional intelligence and ethical considerations of these models while addressing privacy and context-specific nuances.

3 Baseline Models

In this work, we compare our proposed models to two established baseline models: T5-small and BART-base. These models were chosen due to their strong performance in various natural language processing (NLP) tasks, including dialogue generation.

3.1 T5-small Model

T5 is a text-to-text transformer model that frames every NLP task as a text-to-text problem, making it particularly suited for next utterance prediction. For this experiment, we utilized the T5-small variant, which is smaller and more efficient than the full T5 model. To align with T5's expected input-output structure, we formatted the input as: "predict next

utterance: <context>." Beam search was employed during output generation to explore multiple possible continuations, and the model was evaluated using BLEU and BERTScore metrics.

3.2 BART-base Model

BART (Bidirectional and Auto-Regressive Transformers) is a pre-trained model designed as a denoising autoencoder, which allows it to handle noisy or incomplete dialogue inputs effectively. We fine-tuned the BART-base model using a sequence-to-sequence approach and utilized the BartTokenizer for preprocessing the input data. The AdamW optimizer was applied, and the best model was selected based on validation loss. Similar to T5, we evaluated BART's performance using BLEU and BERTScore.

3.3 Baseline Performance

The performance of both baseline models on key evaluation metrics is summarized in the following table:

Table 1: Baseline Model Performance

Model	BLEU	BERT Precision	BERT Recall	BERT F1
T5	0.0186	0.8611	0.8443	0.8524
BART	0.0219	0.8648	0.8488	0.8564

3.3.1 Graphical Representation of Model Performance

T5-small Model Performance: The following graph illustrates the performance of the T5-small model on BLEU and BERTScore:

Figure 1: T5-small Model Performance val and train loss.

BART-base Model Performance: The following graph illustrates the performance of the BART-base model on BLEU and BERTScore:

Figure 2: BART-base Model Performance on train and val loss.

Both T5 and BART perform competitively in terms of BERTScore, and BART slightly outperforms T5 inaluation metrics. However, both models still exhibit low BLEU scores, suggesting that while the models achieve semantic alignment, they struggle with surface-level similarity and exact n-gram matches, which is common in dialogue generation tasks.

4 Methodology

4.1 Model 1

4.1.1 Sentiment Infused Bart Model

To incorporate emotional understanding into the generation process, we propose a novel model that extends the standard BART encoder-decoder architecture by incorporating external sentiment features derived from a pretrained sentiment analysis model. This approach is particularly suited for therapeutic dialogue generation, where emotional context plays a critical role in shaping appropriate and empathetic responses.

4.1.2 Data Preprocessing

We first preprocess the mental health counseling dataset by segmenting long conversations into multiple samples. Each sample contains up to six recent utterances, ensuring that the final utterance in the segment is a therapist's statement. If the original conversation ends with a person's utterance, we treat the last therapist utterance before it as the target. Redundant input segments are removed via deduplication.

4.1.3 Sentiment Feature Extraction

We utilize the `cardiffnlp/twitter-roberta-base-sentiment` model to extract sentiment embeddings from the

input text. Only the `[CLS]` token output is used to represent the overall sentiment of the input. The sentiment model is frozen during training to retain its pretrained knowledge.

4.1.4 Fusion Module

A multi-head attention-based `SentimentFusion` module integrates token embeddings from the BART encoder with sentiment embeddings. The fusion process consists of:

- Concatenating the token embeddings with the sentiment embedding (broadcasted across the sequence).
- Projecting the combined embedding back to the original hidden dimension.
- Applying multi-head self-attention to the fused representation.
- Adding a residual connection from the original token embedding.

4.1.5 Conditional Generation

The fused embeddings are passed to the BART encoder as `inputs_embeds`, bypassing the default token embedding lookup. The BART decoder then generates the target therapist response using teacher forcing during training and top-p sampling with beam search during inference.

4.1.6 Training Details

The model is trained using the following configuration:

- **Loss Function:** Cross-entropy loss on decoder outputs.
- **Optimizer:** AdamW with learning rate $1e^{-5}$.
- **Batch Size:** 8
- **Epochs:** 5

4.1.7 Evaluation Metrics and Results

The performance of the `SentimentInfusedBart` model was evaluated using standard metrics for natural language generation, which measure both lexical and semantic similarities between generated responses and ground-truth references. The results are summarized in Table 3.

Table 2: Evaluation Results for SentimentInfusedBart Model

Metric	Score
BLEU (Average)	0.0104
BERTScore (F1)	0.8480
ROUGE-1 (F1)	0.1428
ROUGE-2 (F1)	0.0187
ROUGE-L (F1)	0.1170

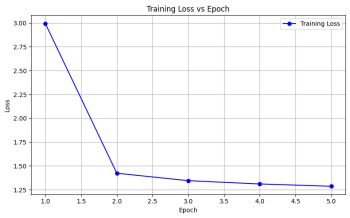

Figure 3: Visualization of Model Performance Across Epochs

4.2 Model 2: Therapeutic Generator-Simulator Model

In this work, we propose a dual-model approach for next-utterance prediction in mental health counseling dialogues. By integrating both supervised learning and feedback learning, the model is trained to optimize for both correctness in dialogue progression and sensitivity to the emotional dynamics of therapeutic conversations.

The proposed framework consists of two main components:

- **Generator Model:** A transformer-based model that predicts immediate therapist responses. It is trained using a standard sequence-to-sequence loss to ensure that the generated responses are fluent and contextually appropriate.
- **Simulator Model:** A complementary transformer model that simulates potential follow-up patient responses based on the generator's output. This model is leveraged to evaluate the emotional impact of the generated responses by integrating an external emotion classifier.

The supervised learning component ensures that the model produces linguistically coherent responses, while the feedback learning component uses an emotion-based reward signal to fine-tune the empathetic quality of the outputs. This dual approach is designed to address the unique challenges in mental health counseling, where both the content and emotional tone of the conversation are crucial for effective support.

4.2.1 Data Preprocessing and Triplet Construction

In our approach, we convert raw counseling dialogues into structured triplets. These triplets serve two purposes: they enable the generator model to

learn immediate response prediction, and they provide the simulator model with a basis for predicting follow-up dialogue.

4.2.2 Triplet Creation

Purpose: In a counseling conversation, context builds progressively as dialogue unfolds. To model this, we create training examples in the form of triplets:

- **History:** Represents the sequence of dialogue turns leading up to a particular point. This preserves the conversation's context.
- **Target Utterance (ut):** This is the immediate response following the history. It acts as the ground truth that the generator model learns to predict.
- **Simulated Next Input (ut1):** Derived from a shifted version of the dialogue, this captures what might naturally follow. It is used by the simulator model.

4.2.3 Methodology

Our approach is based on a dual-model architecture with a hybrid training objective, and consists of three main components:

- **Data Preprocessing and Triplet Construction:** Raw counseling dialogues are processed into triplets—comprising the dialogue history, target utterance, and a simulated next input—to capture conversational context and natural progression.
- **Dual-Model Architecture:** Two transformer-based models (initialized from T5-small) are used:
  - A **Generator Model** that predicts immediate responses using sequence-to-sequence learning.
  - A **Simulator Model** that generates potential follow-up responses, aiding in evaluating and refining the generated outputs.
- **Hybrid Training Objective:** The generator is trained with a weighted loss function combining:
  - **Supervised Learning Loss:** Cross-entropy loss calculated against the true responses.
  - **Feedback Learning Loss:** A reward signal derived from an emotion classifier applied to simulated follow-up responses, enhancing the model's empathetic quality.

This integrated strategy enables the model to produce context-aware and emotionally supportive dialogue responses.

4.2.4 Evaluation Metrics and Results

Figure 4: Visualization of Model Performance Across Epochs

The BLEU score is a widely used metric in natural language processing that measures the n-gram overlap between the generated response and reference utterances. For our model, the BLEU score is 0.0253.

Although this value appears low, it is important to note that BLEU is known to be less effective for dialogue generation tasks, where there can be a wide variety of valid responses, leading to lower n-gram match rates.

4.2.5 BERT-based Evaluation Metrics

To overcome the limitations of BLEU in the context of dialogue, we also employed BERT-based metrics that capture the underlying semantic similarity between generated responses and reference texts. The results obtained are as follows:

- **BERT F1 Score:** 0.8151
- **BERT Precision Score:** 0.8154.
- **BERT Recall Score:** 0.8157

The results are summarized in Table 3.

Table 3: Evaluation Results for T-GenSim Model

Metric	Score
BLEU Score	0.0253
BERT F1 Score:	0.8151
BERT Precision Score	0.8154
BERT recall Score:	0.8157

5 Dataset

About the Dataset. The dataset comprises **mental health counseling dialogues**, curated and structured for the task of **next utterance prediction**. Each dialogue is segmented into sequences of utterances from therapy sessions, with the goal of predicting the next expected utterance.

The dataset is divided into:

- **4008 training samples**
- **576 validation samples**
- **968 test samples**

Each sample in the dataset contains:

- **Input:** A sequence of prior utterances in a therapy session, separated by the special token `[SEP]`.
- **Target:** The next expected utterance in the conversation.

Example. Input:

```
T: Hi you how to do it today? [SEP] P: Great.  How are you? [SEP] T: I'm doing well. Thanks for asking.
```

Target:

```
"So you're doing great."
```

6 Results and Findings

Overall, both models exhibit challenges related to the inherently open-ended nature of therapeutic dialogues. While standard language-generation metrics (e.g., BLEU) often appear low in such tasks, BERT-based semantic metrics suggest a more favorable alignment with reference responses. This emphasizes the importance of using diverse evaluation metrics to capture the richness and complexity of mental health counseling conversations.

6.1 Key Takeaways

- **Emotion-Aware Generation:** Incorporating sentiment analysis modules can improve the empathetic quality of outputs.
- **Semantic Matching vs. Exact Overlap:** Metrics focusing on semantic alignment (e.g., BERTScore) may be more representative for counseling dialogues than pure n-gram overlaps.
- **Potential of Hybrid Architectures:** Combining standard supervised learning with feedback learning (via a simulator) shows promise in balancing linguistic fluency with empathy.

7 Conclusion

The experimental findings underscore the potential of transformer-based approaches to handle next utterance generation in mental health counseling dialogues. Although exact n-gram overlaps remain low, semantic similarity and emotional appropriateness are better reflected in higher BERT-based metrics. Future work could explore additional specialized losses, domain adaptation, and user-centric evaluation criteria to further refine empathetic and contextually relevant therapy responses.

8 Experimental Setup

8.1 Objective

The goal of this experiment was to evaluate two different models for generating therapeutic responses. The first model, the **Therapeutic Generator-Simulator Model**, is designed to simulate therapeutic conversations. The second model, the **Sentiment Infused Model**, integrates sentiment analysis to infuse emotional context into the responses.

8.2 Model 1: Therapeutic Generator-Simulator Model

The **Therapeutic Generator-Simulator Model** is designed to create therapeutic responses based on given input prompts. It was trained on a dataset of therapeutic conversations, ensuring that the generated outputs are empathetic and contextually appropriate. The model uses a transformer-based architecture to understand and generate text in a coherent and human-like manner.

Training Details:

- **Architecture:** Transformer-based, designed for text generation.
- **Dataset:** Therapeutic dialogue dataset.
- **Optimization:** Adam optimizer with a learning rate of 0.0001.
- **Epochs:** 5 epochs.
- **Batch Size:** 8
- **Loss Function:** Cross-entropy loss.

8.3 Model 2: Sentiment Infused Model

The **Sentiment Infused Model** is an enhancement of the basic transformer-based model. It includes an additional sentiment analysis layer that allows the model to generate responses with emotional context. This model is intended to improve the quality of interactions by infusing sentiment and empathy into the responses, making it more suitable for therapeutic applications.

Training Details:

- **Architecture:** Transformer-based with sentiment analysis integration.
- **Dataset:** Combination of therapeutic dialogues and sentiment-labeled data.
- **Optimization:** Adam optimizer with a learning rate of 0.0003.
- **Epochs:** 3 epochs.
- **Batch Size:** 16
- **Loss Function:** Cross-entropy loss.

8.4 Evaluation Metrics

The following evaluation metrics were used to assess the performance of the models:

- **BLEU Score:** Measures the n-gram overlap between the generated responses and reference responses.
- **BERTScore (F1):** Measures the semantic similarity between the generated and reference responses.
- **ROUGE Scores (1, 2, L):** Evaluates the recall of n-grams between the generated responses and reference responses.

8.5 Hardware and Software Specifications

Hardware:

- **GPU:** NVIDIA Tesla V100.
- **CPU:** Intel Xeon Gold 6240.
- **RAM:** 64GB.

Software:

- **Framework:** PyTorch for model training.
- **Libraries:** Transformers, NumPy, Pandas, Scikit-learn.
- **Python Version:** 3.8.

9 Future Work

Future improvements to these models can include the following key directions:

- **Reinforcement Learning with Human Feedback (RLHF):** Explore the use of reinforcement learning techniques to allow models to learn from real-world user feedback, leading to more personalized and context-aware responses.
- **Multimodal Learning:** Incorporate multimodal data, such as voice tone or facial expressions, alongside text input. This could help in modeling empathy more effectively, leading to more emotionally aware and responsive outputs.
- **Dynamic Context Tracking:** Develop mechanisms for tracking long-term conversation context. This would allow the model to maintain emotional coherence over extended interactions, crucial for therapeutic applications where continuity is key.
- **Real-world Evaluation:** Conduct evaluations with clinical experts or psychologists to assess the models' efficacy in real-world therapeutic settings. Their feedback will be invaluable in refining the models and ensuring they meet practical therapeutic needs.

References

Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. 2024. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. *arXiv preprint arXiv:2402.12738*.

Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023*.