Offline Chat-Reply Recommendation System – GPT-2

Objective

To build an offline chat-reply recommendation system using Transformer-based models that can predict the next possible reply from User A when User B sends a message, leveraging previous conversation history for context-aware response generation.

Dataset Overview

Two-person conversation data was provided in the file *conversationfile.xlsx - userAuserB.csv*, containing 22 messages across multiple dialogues. Each record includes Conversation ID, Timestamp, Sender (User A or User B), and Message text. Preprocessing paired each User B message with the subsequent User A reply, forming context–response training pairs for fine-tuning.

Data Preprocessing

- Loaded and sorted messages chronologically per conversation ID.
- Removed unnecessary characters, quotes, and extra whitespace.
- Formed context–reply pairs (User B message and next User A response).
- Included short conversational history for improved context learning.
- Saved cleaned data as processed_conversations.csv.

Model Choice & Justification

Among the allowed models (BERT, GPT-2, and T5), GPT-2 was chosen for its autoregressive generation capability, which makes it ideal for reply prediction in dialogue. Unlike BERT (which is bidirectional and non-generative), GPT-2 natively supports sequence continuation given a context, and can be fine-tuned offline without requiring internet connectivity. The model uses causal language modeling loss with context masking so that only reply tokens contribute to optimization.

Training Setup

Fine-tuning was conducted offline using preloaded GPT-2 weights. The dataset was split 80/20 into training and validation sets. Special tokens such as , , , and were introduced to differentiate dialogue roles. The model was trained for 3 epochs using AdamW optimizer with a learning rate of 5e-5 and gradient accumulation to support limited hardware. Evaluation metrics included Perplexity, BLEU, and ROUGE-L.

Evaluation Metrics

- Perplexity – measures fluency and confidence of the model.
- BLEU – compares generated reply n-grams with reference replies.

- ROUGE-L – measures longest common subsequence overlap with references.

Results Summary

The model achieved low validation loss and reasonable BLEU/ROUGE scores given the small dataset. Generated replies were contextually coherent, reflecting local conversational flow. Example results (from samples.jsonl) show GPT-2 effectively mirroring User A's tone and continuity.

Offline Deployment Feasibility

The model and tokenizer were exported locally as Model.joblib and artifact directories. During offline inference, the generate_reply(context) function can be invoked to produce context-aware replies from User A. Since GPT-2 weights are preloaded, no internet dependency exists during runtime, satisfying offline deployment constraints.

Conclusion

This offline GPT-2–based chat-reply recommendation system successfully processes conversational data, fine-tunes a generative model, and evaluates outputs using standard NLP metrics. It demonstrates coherent reply generation and efficient offline inference, aligning with the Round 4 AI/ML Developer Intern objectives.