A Brief Report

Advanced Generative Design of a Chatbot using Cornell Movie Dialogs Corpus

INTRODUCTION

This project presents the design and implementation of an advanced generative chatbot using the Cornell Movie Dialogs Corpus. The chatbot employs a seq2seq model within conversation generation, focusing on meaningful and contextually appropriate responses.

CHALLENGES FACED AND SOLUTIONS IMPLEMENTED

Data Issues: Large datasets with missing entries and inconsistencies posed challenges during preprocessing. We addressed this by doing very extensive EDA and data cleaning.

-Text Preprocessing: Tokenizing and normalizing dialogs were quite involved but resolved using NLTK and BERT tokenizer

-Model Training: The model fine-tuned a transformer-based seq2seq to get contextually correct answers.

3.Model Architecture and Rationale

The model selected is a transformer-based seq2seq model with an attention mechanism to better handle context. This enables the model to pay attention to specific parts of the input conversation while generating responses, which should make it more coherent.

4.Evaluation Results and User Feedback

Basic BLEU scores for the model are somewhat acceptable. Users' preprocessed feedback says responses are generally good but, in some instances, tolerate out-of-context.

5. Future Developments and Scalability Options

Fine-tune the model with large-scale datasets and use GPT-based models. Scale the chatbot by cloud deployment via AWS/GCP and develop a userfriendly web interface to enhance interaction.

6. Conclusion

The project will unleash the potential that seq2seq models can offer while building chatbots with improvements to be made in coherence and response accuracy, so it is good to go for enhancement and scaling.Advanced Generative Chatbot Design using Cornell Movie Dialogs Corpus