Suko - The Chatbot

*Ishwar Babu

Abstract—**This project involves the development of a sentiment-aware chatbot, named Suko, capable of engaging users in natural language conversations. Leveraging datasets from movie scripts and metadata, the chatbot utilizes sentiment analysis and natural language processing (NLP) techniques to provide contextually relevant responses. Key components of the project include sentiment analysis using the VADER Sentiment Analyzer, question-answer retrieval from movie dialogue datasets, and generating responses with GPT-2. The chatbot also offers specific functionalities such as providing movie and character information based on user queries. By integrating BERT for sentiment analysis and GPT-2 for response generation, Suko aims to enhance user interaction by understanding and responding to the emotional tone of the conversations. The project demonstrates the application of state-of-the-art NLP models in creating an intelligent conversational agent with diverse capabilities.**

Index Words: GPT-2, natural language processing, data modeling, Vader Lexicon, transformer

I. Introduction

In the rapidly evolving field of artificial intelligence (AI), chatbots have emerged as a prominent application of natural language processing (NLP) and machine learning (ML). These conversational agents are designed to simulate human-like interactions, providing users with seamless and engaging experiences across various domains such as customer service, personal assistance, and entertainment. This project presents the development of Suko, a sentiment-aware chatbot that leverages advanced NLP techniques and datasets from movie scripts to interact with users in a meaningful and contextually relevant manner.

Suko is designed to analyze user input for sentiment, enabling it to understand and respond to the emotional tone of the conversation. This feature is achieved through the integration of the VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analyzer, which classifies user messages as positive, negative, or neutral. By incorporating sentiment analysis, Suko can tailor its responses to better align with the user's emotional state, enhancing the overall user experience.

In addition to sentiment analysis, Suko utilizes a rich dataset of movie dialogues and metadata to provide informative and entertaining responses. The dataset includes movie conversations, lines, titles, and character information, allowing Suko to retrieve specific details about movies and characters upon request. This capability is further augmented by the use of pre-trained NLP models such as BERT (Bidirectional Encoder Representations from Transformers) for sentiment classification and GPT-2 (Generative Pre-trained Transformer 2) for generating human-like responses.

II. Literature Review

The development of chatbots has been a significant area of research within the field of artificial intelligence and natural language processing. The evolution of chatbots can be traced from rule-based systems to the more sophisticated, machine learning-driven models used today. This literature review highlights key advancements and methodologies in chatbot development, sentiment analysis, and the application of pre-trained language models.

A. Evolution of Chatbots

Early chatbots, such as ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975), were rule-based systems that relied on pattern matching and predefined scripts to simulate conversation. These systems were limited by their lack of understanding and inability to adapt to new contexts. The introduction of statistical methods and machine learning algorithms marked a significant shift in chatbot capabilities, allowing for more dynamic and context-aware interactions.

B. Sentiment Analysis in Chatbots

Sentiment analysis is a crucial component for enhancing user interaction by enabling chatbots to understand and respond to the emotional tone of user inputs. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a popular sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto Gilbert, 2014). It uses a combination of lexical heuristics and a predefined sentiment lexicon to classify text as positive, negative, or neutral. The integration of sentiment analysis in chatbots can improve user satisfaction by making interactions more empathetic and contextually appropriate (Poria et al., 2017).

C. Pre-trained Language Models

The advent of pre-trained language models such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) has revolutionized natural language processing tasks, including chatbot development. BERT, which stands for Bidirectional Encoder Representations from Transformers, provides deep bidirectional representations by conditioning on both left and right context in all layers. It has set new benchmarks in several NLP tasks, including sentiment analysis and question answering.

GPT-2, developed by OpenAI, is a generative model capable of producing coherent and contextually relevant text. It utilizes a transformer architecture and is trained on a diverse dataset, enabling it to generate high-quality responses in a conversational setting. The combination of BERT for sentiment analysis and GPT-2 for response generation has shown promising results in creating more intelligent and interactive chatbots (Wolf et al., 2020).

D. Challenges and Future Directions

Despite significant advancements, several challenges remain in chatbot development. These include handling ambiguous or out-of-scope queries, maintaining coherent context over long conversations, and ensuring ethical considerations in response generation. Future research is likely to focus on improving context management, integrating multimodal inputs, and developing more robust evaluation metrics for chatbot performance.

In summary, the literature indicates that the integration of sentiment analysis and pre-trained language models has significantly enhanced chatbot capabilities. By leveraging movie datasets, chatbots like Suko can provide engaging and informative interactions. The ongoing advancements in NLP and machine learning hold promise for further improving the functionality and user experience of conversational agents.

III. Dataset

The dataset used for this project includes several components derived from the Cornell Movie Dialogues Corpus. This corpus is a rich collection of movie conversations, lines, and metadata, which provides a diverse and natural language resource for training and evaluating dialogue systems.

A. Movie Conversations

The file contains the conversation structure in movies. Each line in this file represents a conversation between characters in a movie

- Character IDs involved in the conversation
- Movie ID
- List of utterance IDs

This file helps in understanding the dialogue flow between characters, which is essential for creating a realistic conversational agent.

B. Movie Lines

The file contains individual lines of dialogue from movies.

- Line ID
- Character ID
- Movie ID
- Character name
- Dialogue text

This file provides the raw text of movie dialogues, which is used to build the question-answer pairs for the chatbot.

C. Movie Metadata

The file includes metadata about the movies.

- Movie ID
- Movie description

This metadata is utilized to provide movie-related information when users ask for it.

D. Character Metadata

The file contains metadata about the characters.

- Character ID
- Character name
- Movie ID
- Character gender

This information helps in answering user queries about specific characters and their roles in the movies.

Link to Dataset: Cornell Movie-Dialog Corpus

IV. Methodology

The methodology section outlines the steps taken to develop Suko, the sentiment-aware movie chatbot, from data preprocessing to model implementation and integration. The process involves several key stages, including data preparation, model selection, training, and the creation of a user interface.

A. Data Preparation

1) Loading and Preprocessing Data:

- **Movie Conversations**: The *movie_conversations.txt* file was read to extract the sequence of dialogue exchanges between characters.
- **Movie Lines**: The *movie_lines.txt* file was processed to map each line ID to the actual dialogue text. This allowed us to pair questions with their corresponding answers.
- **Question-Answer Pairs**: Using the conversation data, question-answer pairs were generated by linking consecutive lines of dialogue. These pairs were stored in a pandas DataFrame for easy manipulation.
- **Movie and Character Metadata**: The *movie_titles_metadata.txt* and *movie_characters_metadata.txt* files were loaded to create dictionaries that map movie and character IDs to their respective metadata, such as titles, descriptions, and character names.

2) Data Structuring: The processed data was structured into a format suitable for training and interaction, including a DataFrame for question-answer pairs and dictionaries for movie and character metadata.

B. Sentiment Analysis

1) Sentiment Analysis Model:

- **VADER Sentiment Analysis**: The VADER (Valence Aware Dictionary and sEntiment Reasoner) tool was utilized to perform sentiment analysis on user inputs. This tool is effective for analyzing sentiments in social media and conversational text.
- **BERT Sentiment Classifier**: A fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model was employed to classify the sentiment of user inputs into three categories: negative, neutral, and positive. The BERT model was loaded using the *transformers* library.

2) Sentiment Integration:

- The sentiment scores were calculated for each user input, and the compound score was used to determine the overall sentiment (positive, negative, or neutral).
- These sentiment scores were displayed to the user along with the chatbot's response, providing transparency and context for the conversation.

C. Dialogue Generation

1) Dialogue Response Generation:

- **GPT-2 Model**: The GPT-2 (Generative Pre-trained Transformer 2) model was used to generate responses to user queries. This model was fine-tuned to produce coherent and contextually relevant dialogue based on the input text.
- The GPT-2 model was implemented using the *transformers* library, and it generated responses by encoding user inputs and producing text outputs.

2) Question-Answer Matching:

- For specific questions about movies and characters, the chatbot searched the preprocessed question-answer pairs for exact or partial matches to provide accurate responses.
- If a match was found, the corresponding answer was retrieved and presented to the user. If no match was found, a fallback response was generated using the GPT-2 model.

D. Movie and Character Information Retrieval

1) Movie Information: When users requested movie information, the chatbot searched the movie metadata dictionary for the requested title. If found, the movie title and description were provided.

2) Character Information: For character information, the chatbot searched the character metadata dictionary for the requested character name. If found, the character's name, the movie they appeared in, and their gender were provided.

E. User Interface

1) Interactive Menu:

- The chatbot featured an interactive menu that guided users through various options, such as asking a question, receiving a fun fact, engaging in small talk, or requesting movie and character information.
- This menu reappeared after each response, ensuring a smooth and user-friendly experience.

2) Conversation Flow: The conversation flow was managed using a simple state machine that tracked the user's current state (e.g., asking a question, requesting movie information). Based on the user input and the current state, appropriate responses and prompts were generated to maintain a coherent dialogue.

F. Implementation

1) Programming Environment: The chatbot was implemented in Python, leveraging libraries such as *pandas* for data manipulation, *transformers* for model loading and inference, and *nltk* for text processing. The interactive user interface was created using a simple loop that took user inputs and displayed the chatbot's responses.

2) Model Loading and Initialization: Pre-trained BERT and GPT-2 models were loaded and initialized using the *transformers* library. These models were fine-tuned on the conversational dataset to enhance their performance in dialogue generation and sentiment analysis.

3) Response Generation: User inputs were analyzed for sentiment, matched against the question-answer pairs, and processed through the GPT-2 model to generate appropriate responses. The chatbot's responses included sentiment scores and contextually relevant information, ensuring a comprehensive interaction.

By following this methodology, Suko was developed to be a robust and interactive movie chatbot capable of engaging users in meaningful conversations while providing accurate movie and character information.

V. Sample Interaction

- **User Input:** *hello*
  - **Sentiment Analysis:** Positive
    Scores: {'neg': 0.222152, 'neu': 0.4893153, 'pos': 0.28853264, 'compound': 0.4893153}
  - **Response:** Hello! How can I assist you today?
- **User Input:** *4*
  - **Sentiment Analysis:** Positive
    Scores: {'neg': 0.20273842, 'neu': 0.48758754, 'pos': 0.309674, 'compound': 0.48758754}
  - **Response:** Please enter the movie name or a question about the movie.
- **User Input:** *The Matrix*
  - **Sentiment Analysis:** Positive
    Scores: {'neg': 0.23579207, 'neu': 0.48001206, 'pos': 0.28419593, 'compound': 0.48001206}
  - **Response:** Movie: the matrix
    Description: 1999
- **User Input:** *2*
  - **Sentiment Analysis:** Positive
    Scores: {'neg': 0.22891816, 'neu': 0.47678962, 'pos': 0.29429224, 'compound': 0.47678962}
  - **Response:** Did you know that honey never spoils?
- **User Input:** *6*
  - **Sentiment Analysis:** Positive
    Scores: {'neg': 0.19520536, 'neu': 0.4836319, 'pos': 0.32116273, 'compound': 0.4836319}
  - **Response:** Goodbye! Have a great day!
- **User Input:** *exit*
  - **Response:** Goodbye! Have a great day!

VI. Future Work

The current implementation of the chatbot demonstrates effective integration of sentiment analysis, natural language processing models (BERT and GPT-2), and data retrieval techniques for providing informative responses. However, there are several avenues for future enhancements and research:

- **Enhanced Contextual Understanding:** Incorporate contextual understanding to improve response relevance, especially in handling ambiguous queries or complex user intents.
- **Multi-turn Dialogue Handling:** Develop capabilities for maintaining context across multiple user interactions to facilitate more natural and coherent conversations.
- **Personalization and User Modeling:** Implement mechanisms for user profiling and personalized responses based on historical interactions and user preferences.
- **Integration of Additional Models:** Explore integration of more advanced language models or domain-specific models to enhance the chatbot's knowledge base and response accuracy.
- **Evaluation and User Feedback:** Conduct comprehensive user studies to evaluate the effectiveness and user satisfaction with the chatbot, incorporating feedback loops for continuous improvement.
- **Deployment and Scalability:** Address deployment challenges and scalability issues to support large-scale deployment across different platforms and environments.
- **Ethical Considerations:** Investigate ethical implications of AI-driven chatbot interactions, including privacy concerns, bias mitigation, and responsible AI practices.

These future directions aim to advance the capabilities of the chatbot, making it more intelligent, responsive, and user-friendly for diverse applications and user scenarios.

VII. Conclusion

In this project, we developed Suko, a sentiment-aware chatbot that integrates advanced natural language processing techniques with sentiment analysis and data retrieval from movie-related datasets. The chatbot successfully demonstrates the capability to engage users in meaningful conversations, provide movie information, and offer fun facts based on user interactions.

Through the implementation of sentiment analysis using VADER and deep learning models such as BERT and GPT-2, Suko effectively assesses the sentiment of user inputs and generates appropriate responses. The integration of these models allows Suko to understand user queries, retrieve relevant movie and character information, and engage in small talk or provide interesting facts.

The development process involved preprocessing and integrating movie-related datasets, including conversations, lines, movie metadata, and character metadata. This structured approach enabled Suko to access comprehensive information about movies and characters, enriching the user experience with relevant and accurate responses.

While Suko demonstrates robust functionality, there are opportunities for further enhancement and research, as outlined in the future work section. These include improving contextual understanding, implementing multi-turn dialogue handling, enhancing personalization, and evaluating the chatbot's performance through user studies and feedback mechanisms.

Overall, Suko represents a significant step towards creating intelligent and interactive chatbot systems that can engage users effectively while leveraging advanced AI techniques for enhanced functionality and user satisfaction in diverse applications.