INTERNSHIP PROJECT DOCUMENT II

ON

FAQs Chatbot using Natural Language Processing (Text Generation) – Understanding Data and Provides Answers

Submitted by

Swarupa Balaji

(Artificial Intelligence Intern)

Submitted to:

Founder- Kanduri Abhinay

CTO - Saniya Begam

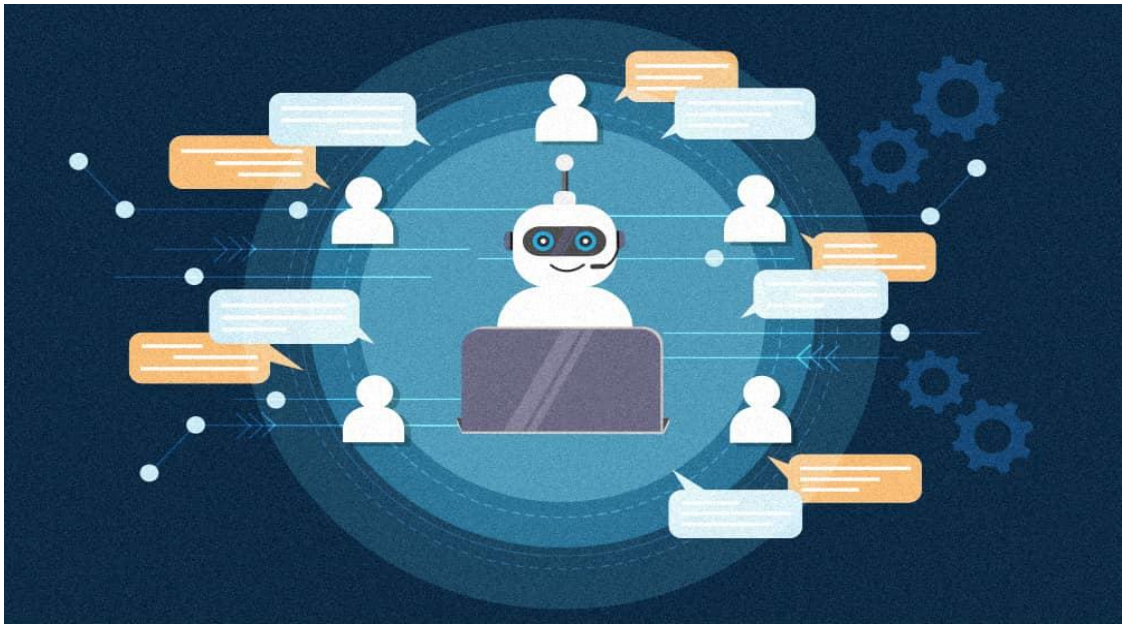
Team Lead – Surya Poojavi

PROJECT II

Title: FAQs Chatbot

Project Description:

Frequently asked questions are a prominent part of a website. Some web apps contain FAQs that answer as many as 5 to 10 questions. Still some customers may have questions that are not in the FAQs. Here an AI chatbot is applicable that would answer user queries using the data it is trained on. I have used an NLP based model that can generate text based on the comma separated data provided to it (A series of questions and answers). The generative AI model BERT of hugging face is frequently used for this use case.



Software and Tools Used in Project:

Python: Widely used for machine learning projects due to its simplicity and extensive libraries can be installed very easily.

Streamlit: A type of open sourced framework that can be used to create website applications' GUI, helps design attractive styles and provides modern templates that can be used to make LLMs.

DistilBERT: It is a lightweight transformer model that is smaller but faster and cheaper model than other LLMs. Mainly, DistilBERT can be used for text generation and optimization.

Packages and Libraries Installed:

Data Loading and Preprocessing:

Pandas: For data manipulation and analysis, allowing easy handling of datasets in tabular form. Some techniques include: renaming columns, handling missing data, converting data formats and dtypes, loading data, grouping categories etc.

Model Development:

Scikit-Learn: Scikit-learn is an open-source library for machine learning, providing efficient tools for classification, regression, clustering, and preprocessing, along with utilities for model evaluation and selection.

In this project, Scikit-learn has been used for:

- Model Selection
- Analysis of accuracy
- Model metrics

Torch: PyTorch is a python library that is exclusively used for creating deep neural networks for different models. For DistilBERT, Torch provides tensor operations to effectively compute multi-dimensional arrays, support GPU acceleration and can be used for integration with other frameworks.

Transformer: This library helps import all the models from generative AI. We can receive tokenizers that help process our data into tokens or data chunks that the model can read easily and train on it. The DistilBERT tokenizer trims the data into tokens and the model trains itself on these tokens.

Cosine-similarity: It is a metric used to measure how similar two non-zero vectors are, regardless of their magnitude. It is particularly used in natural language processing (NLP) and information retrieval, to compare the similarity between text, or other types of data.

Algorithm used in Model:

DistilBERT is a smaller, faster, and more efficient version of the BERT (Bidirectional Encoder Representations from Transformers) model. Developed by Hugging Face, it retains much of BERT's performance while significantly reducing the model size and inference time. The primary motivation behind DistilBERT is to make transformer models more accessible for deployment in resource-constrained environments or applications that require real-time processing.

Knowledge Distillation:

DistilBERT is a student model that learns to predict the output of the parent model BERT. Firstly, BERT is pre-trained with the data available and then transfer of knowledge happens from the pre-trained teacher model (BERT) to the student model (DistilBERT). This is called **Knowledge Distillation**.

During training, the student model is exposed to the same input as the teacher model, and it learns to predict the teacher's output. The student model's loss function combines the traditional cross-entropy loss for the actual labels and the distillation loss, which minimizes the difference between the teacher's and student's outputs.

DistilBERT produces logits that are converted to probability and then the model is fine-tuned.

Applications of DistilBERT:

DistilBERT is used in various applications, including but not limited to:

- Text classification (e.g., sentiment analysis)
- Named entity recognition (NER)
- Question answering systems
- Text summarization

Project Execution:

- Data in the form of csv file (consists of question and answers) is imported using pandas.

Question	Answer
Hi or Hello	Hello! How can I assist you today?
What are you/Who are you	I am a virtual AI assistant, here to assist you!
How were you trained?	I was trained using DistilBERT transformer model
How can I create an account?	To create an account, click on the 'Sign Up' button on the top right corner of our website and follow the instructions to complete the registration process.
What payment methods do you accept?	We accept major credit cards, debit cards, and PayPal as payment methods for online orders.
How can I track my order?	You can track your order by logging into your account and navigating to the 'Order History' section. There, you will find the tracking information for your shipment.
What is your return policy?	Our return policy allows you to return products within 30 days of purchase for a full refund, provided they are in their original condition and packaging. Please refer to our Returns page for detailed instructions.

- The tokenizer and the model from DistilBERT are defined where the questions are embedded into tokens. Similar datapoints are put together in batches using cosine similarity and matched to the respective answers.
- Then the user input is mapped to the model, which runs through its logic to find the best answer to the question and sends it to the application.

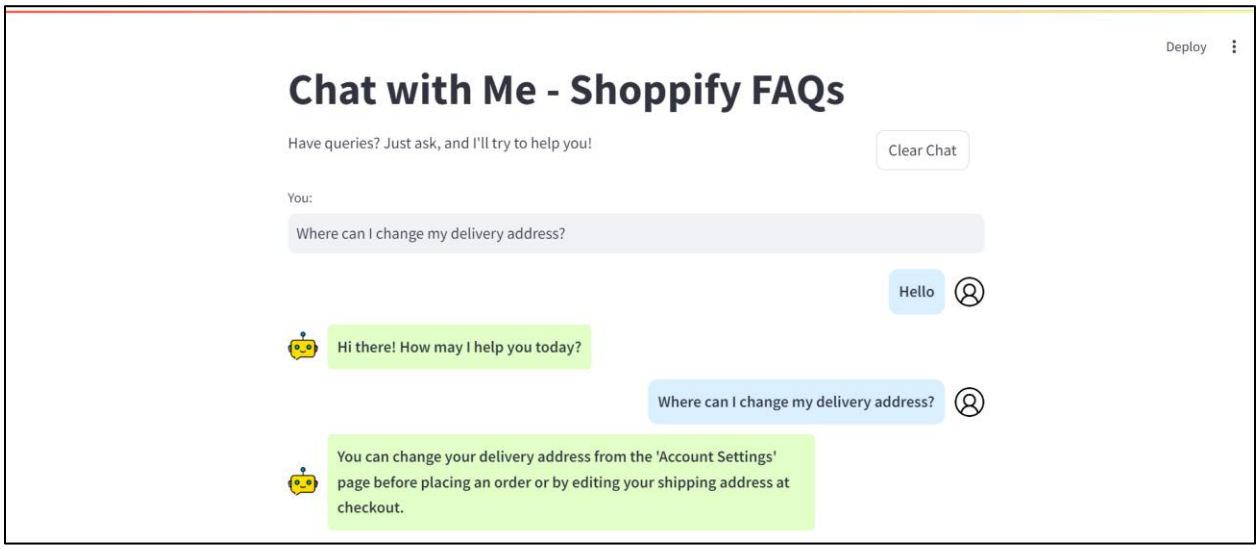
- Using the testing data, the model is tested to see its accuracy. A confusion matrix helps in finding times the model did correct or incorrect predictions with respect to the true output and predicted output.
- Accuracy score, precision and recall used to evaluate the model.

Libraries used:

- Streamlit for web application UI.
- Pandas for loading the csv data.
- DistiBERTTokenizer and DistilBERTModel for converting text to tokens and model development.
- Cosine Similarity for metrics (mapping similar datapoints together)
- PyTorch for overall implementation of the model.

APPLICATION OUTPUT:

Chatbot for an online shopping site FAQs:



- Clear chat allows me to clear the conversational chat logs and start a new chat.
- As much as we want, new questions and answers can be added to the data and the model will be automatically trained on this data, to keep the chatbot updated.

Sources:

- Kaggle
- ChatGPT
- Youtube
- Huggingface

- Blackbox.ai