Arun Singh Negi

Machine Learning

arunpycodecmaster@gmail.com 8077840135 Delhi, India LinkedIn Github

PROFILE

Data science professional, specializing in real-time problem-solving in machine learning and deep learning. Proven track record in developing scalable code in a cloud environment. Keen to stay current with industry trends for efficient, impactful data-driven solutions.

SKILLS

statistics — *Descriptive statistics* | *Inferential statistics*, **Language** — *Python,SQL,HTML,CSS,JS*, **Machine Learning** — *Supervised* | *Unsupervised Learning*, **Natural Language Processing-NLP** — *RNN* | *LSTM* | *GRU* | *Word2vec* | *Transformers* | *Encoder-Decoder*, **Generative AI** — *LLM* | *Vector Databases-Pinecone,FAISS,Chroma* | *llama-2,OpenAI,Mistral,Groq,Gemma*, **MLOPS** — *GitHub* | *Docker* | *AWS ECR* | *AWS Bucket* | *AWS EC2* | *CI-CD*, **Data Analytics** — *Pandas, Seaborn, Matplotlib*

PROFESSIONAL EXPERIENCE

Teleperformance May 2022 – Jul 2023
Sr. Data Scientist Gurugram, India

- **Built a RAG chatbot** that answers queries using knowledge base documents and chat context. Integrated VectorDB for retrieval and a transformer for generation.
- Used VectorDB vector search to retrieve relevant troubleshooting documents, improving accuracy by grounding responses in the right context.
- Fine-tuned Transformer on 2 years of tech support chat data for generating natural, conversational responses to common support queries.
- Designed a machine learning pipeline including data ingestion, transformation, model training, evaluation, and deployment for seamless updates.
- Collaborated with teams to deliver the final product, deploying the chatbot and a scalable, continuously learning system.

Teleperformance Jul 2020 – May 2022
Data Scientist Gurugram, India

- Worked on project for **change request classification and prediction of closure notes.**
- Designed an outage Machine learning classification model to categorize change requests and predict closure notes.
- Experimented with many models for classification model.
- Created an ML pipeline with components for data ingestion to model pushing stages.
- Closure note predicted implemented with NLP techniques.
- Implimented CICD pipeline to deploy model on AWS cloud.

PROJECTS

Chatbot
Gen AI
Github:https://github.com/Arun02DS/chatbot_medical.git
-**Tech Stack** - Python,Pinecone vector db,Langchain, Meta lamma2, Flask
-Utilized pre-trained LLAMA2 model to deploy Flask-based application.
-Trained model on medical encyclopedia data, enabling it to answer related queries proficiently.
-Employed text embedding with overlap processing, stored in Pinecone for efficient retrieval.

Text Summarization
Summary of Text (LLM project)
GitHub: https://github.com/Arun02DS/Text-summarization.git

- **Tech Stack**- Python, Huggingface Transformers, Pegasus-cnn_dailymail,Github, FastAPI, Docker, AWS
- I implemented a text summarization solution by incorporating a pre-trained model within an end-to-end pipeline.
- The pipeline encompasses data collection, transformation, training, and evaluation stages.
- An application has been created to facilitate real-time implementation checks, whether offline or on the cloud.
- After the process of dockerization, the image has been successfully deployed to the AWS cloud.

Thyroid Detection Problem
Classification Problem
GitHub: https://github.com/Arun02DS/Thyroid_detection.git

- **Tech Stack** - Python, MongoDB, GitHub Actions, Docker, S3 bucket, AWS EC2, AWS ECR, Apache Airflow.
- I designed an ML pipeline with six key components such as data ingestion,data validation, data transformation, etc. emphasizing robustness through integrated logging and thorough exception handling in the codebase.
- Initiated data ingestion by extracting data from MongoDB and incorporating it into the pipeline
- Created an image using Docker and deployed it to the AWS cloud through GitHub Actions.
- AWS components come into play as the artifacts are stored in an S3 bucket, images are housed in the ECR repository, and the execution of code takes place within an EC2 instance integrated with Apache Airflow.

MCQ Generator
Gen AI
GitHub: https://github.com/Arun02DS/Langchain_generator.git

- **Tech Stack**- Python, Langchain, OpenAI, Streamlit

- An streamlit application was trained on Text data and responce output as MCQ questions based on that text was developed.
- It uses OpenAI "gpt-3.5-turbo" model to build a sequential chain with prompt and input parameters.
- Model get trained on input text using given prompt give responce as a JSON text.

EDUCATION

Masters of Technology — Oct 2013 – Jul 2015
Govind ballabh pant university of agriculture and technology — Pantnagar, Uttrakhand, India

Bachelor of Technology — Aug 2007 – Apr 2011
Uttrakhand technical university — Dehradun,Uttrakhand, India

CERTIFICATES

Oracle Cloud Infrastructure 2023 Certified Data Science Professional | **Full Stack Data Science Bootcamp** —
Machine Learning, LLM, GenAI

DECLARATION

I hereby declare that the details furnished above are true and correct.

Arun Singh Negi