Self-Introduction

My name is Keshava Sharma, and I am a Data Scientist with 4.4 years of relevant experience. My expertise lies in Generative AI, Computer Vision, Machine Learning, and Natural Language Processing (NLP). Over the years, I have worked on various models to solve business problems using a range of algorithms and technologies.

In Generative AI, I have hands-on experience in building a chatbot using Langchain, RAG and Django frameworks to address the business needs of the client.

In the field of Computer Vision, I worked on developing a YOLO model for wound detection at AIIMS, New Delhi.

In Machine Learning, I am proficient in using classification and regression algorithms like Logistic and Linear Regression, Decision Trees, KNN, and ensemble techniques- boosting and Bagging.

Over the years, I have gained experience in handling end-to -end projects, both as an individual and a team player, starting from data collection, preprocessing, analysis, model building, validation, fine tuning using hyperparameters, model deployment, and presenting the results to my supervisors and clients.

GenAI Project

Problem Statement-

After discussion with the client, we gained access to their database containing the documents regarding policies on offers, discounts, refund, pricing, upcoming sales, closing time, opening time, etc.

Workflow-

To simplify this, we created our own master dataset which contained all the necessary details of the client pdf files. Using python library pypdf2, we extracted the text from PDFs and stored it into JSON format. We then converted the extracted text into vector representations using OpenAI embeddings (text embeddings ada-002) (It is the current best, giving small vector sizes and better semantic understanding).

The chunk size was 500 tokens with a chunk overlap of 100 characters.

Then, we connected to GPT 3.5 Turbo model, we performed transfer learning and writing prompts. When a user query is received, relevant chunks are retrieved from the vector database based on semantic similarity search inside the SQL Server and after chunks retrieval, the response is generated by LLM.

By integrating Lang Chain frameworks, we created a smooth system where user queries were matched with the appropriate relevant and accurate responses. This streamlined data handling and delivered smart, context-aware support.

Wound detection Project

Problem statement-

The task was to develop a wound detection model to identify different types of wounds. This model will assist healthcare professionals in tracking the type of wound, wound healing and incidences of infections.

Dataset size- We took photos of 150 patients of Day 1, 3, 7 and 14, so a total of 600 photos. Different classes were made using Labelme to distinguish different types of wounds, their healing and incidences of infections. The initial data size was around 3 GB.

Steps-

1. We collected the data in the form of images from hospital records.
2. We annotated each instance with pixel-level masks, bounding box coordinates and assigned them class labels using labelme. A healthcare professional assisted in the annotation process.
3. We resized the images (640,640) and normalized the images to ensure consistency in the input data.
4. We applied data augmentation techniques like flipping, rotation, colour and brightness adjustments to increase the diversity of the data.
5. We split the data into training, validation and test set (using train-test-split).
6. We trained the YOLOv5 model on the training set.
7. We evaluated the model's performance on validation set using metrics- IoU (Intersection over union) and mAP (mean average precision).
8. We experimented with hyperparameters of the model using learning rate, batch size and anchor-box size to optimize the model's performance.
9. The final testing after fine-tuning the hyperparameters was done on the test set.
10. Deployed the model on AWS cloud using EC2 instance.

Sentiment Analysis Project

Problem statement

This project involved a thorough examination of customer reviews to extract insights regarding product preferences and opportunities for enhancement. By applying sentiment analysis techniques, the project aimed to measure customer satisfaction levels and contribute constructive feedback to the product development process.

Our client wanted to conduct sentiment analysis of retail product and classifying the reviews into positive and negative sentiment from that they want to measure customer satisfaction levels and contribute constructive feedback to the product development process

Steps-

1. The data was handed over by client in csv format.
2. I pre-processed the data by lowercasing, removing HTML tags, URLs, chat word treatment, removing punctuation and stop word removal.
3. Then I did lemmatization to bring the words in their lemma form.
4. I then used the TF-IDF vectorizer to convert words into vectors.
5. Then, we various algorithms for modelling.
6. We achieved highest accuracy using Logistic Regression.
7. Then we deployed the model on Streamlit.