Mayur Joshi

Address: Zolo Fireflies, Pune.

Mob: 7977922167 | Email: mayurjoshi576@gmail.com.

Profile Summary

- Data Scientist with **3 years** of experience in developing end-to-end data science products and delivering client projects.
- Good understanding of Data Pre-processing, Machine Learning Algorithm and Deep Learning Algorithm training and deployments.
- Worked on **Information Retrieval**, **Time Series Forecasting**, **Classification, Sentiment analysis and Clustering** Problems.
- Hands on experience in **NLP** tools, Sentence and Word Embeddings like **BERT, W2V**.
- Hands on Experience in **Sklearn**, **Numpy**, **Pandas**, **Matplotlib**, **Tensorflow** and **Keras** Python libraries.

SKILLS

- **Python**
- **SQL**
- Machine Learning: - **KNN, SVM, Xgboost, Random Forest, Decision Tree, Logistic Regression**.
- Deep Learning: - **LSTM, GRU, CNN, MLP**
- Time Series Forecasting:- **S-ARIMA, Holt-Winters**
- **Scikit Learn, TensorFlow, Pytorch, Keras**
- **AWS Services**
- **Apache Solr**
- Flask, Docker, CI/CD

WORK EXPERIENCE

Machine Learning Engineer

DigitalMain Pune | From June 2019 to Present

As a ML Engineer, I am working on digital assistant platform in an Agile environment.
My responsibilities included: -

- Question Answer retrieval system using **Solr** and **BERT** and **Word Vectors**.
- Build end to end Question generation application from paragraph using **BERT** and **GPT2** with use of **GPU, Docke**r and **Flask** Framework.
- Worked on Intent classification and Key Phrase extraction model.
- Research and Implementation of **PDF Parsing** and Image and Caption extraction techniques.
- Worked on Python Software developments, Model Integrations, handling production deployments.
- Worked on Scripting of Log rotation, Error emails and automated Software code deployments.
- Collaborating with team for designing of system architecture and problem formulation.

Domain Specific Question with Context Extraction from Slack, MS-Teams, Emails, Web and Mobile.

To provide answers for user asked question through **Slack, MS-Teams, Emails, Web and Mobile** designed pipeline for domain specific question and context extraction using Machine Learning, **NER** and rule-based approach.

Roles and Responsibilities: -

1. Data cleaning by analysing different Email formats, Slack, MS-Teams chat conversation and Signature patterns.
2. Intent and question classification of Emails and chats into domain specific and generic categories (Meeting request, Follow Up) using **W2V** and **SVM**.
3. **NER** Model for context keyword extraction and evaluated performance with RAKE/YAKE model.

Search Relevance Architecture design and development.

It includes Semantic based search on Question-Answer (FAQ), PDF extracted paragraph, Historical email threads. Relevance based ranking of search results. Removing or collapsing of Duplicates answers. Autosuggest questions on user typed query based on characters.

Roles and Responsibilities: -

1. **Solr** Core setup for implementation of Full text search on QA, PDF paragraph and vectors.
2. For Semantic search created domain specific **W2V** model using historical email data, crawling domain specific data from websites, pdf.
3. Built relevancy classifier using features like word share, distance based, Fuzzy logic and Word vector based to classify into relevant and irrelevant and re-rank obtained search results using confidence score.
4. Implementation of **Multiword Synonyms** and **Edismax query** parsing in Solr.
5. Key-Phrase Extraction using **LSTM** and **W2V** for boosting Solr results.
6. Research on NLP Pretrained Model like **BERT** for Domain specific and IR use case.
7. Fine-tuned and end to end implementation of **BERT MLM**, **NLI** and **Cross Encoder** Model. Achieved around **89%** search accuracy on top 3 results.

PDF parsing.

Extracting, parsing and re-structuring pdf documents into head, paragraphs and title.

Roles and Responsibilities: -

1. Researching and validation of available open-source pdf parser. Based on comparison GROBID well suit for our use case.
2. Retraining of specific components of model based on custom tag. Created tagged data on business specific pdf. Retraining of Segmentation and Header extraction model to get head, paragraphs and title and to remove unwanted text from pdf like header, footer, caption of images and tables.
3. Worked on creation of API. It takes input as pdf and output as paragraphs with heads and title of pdf.
4. Production ready deployment of solution using Docker and CI/CD on AWS.

Question generation from Paragraph.

The question generation is a transformer-based model based off of **GPT-2** and **BERT** QA. GPT2 used for generation of questions. **BERT** QA used for validating generated question has belong to paragraph or not.

Roles and Responsibilities: -

1. Research and Validation of **GPT2** and **BERT** QA for use case
2. End-End Pipeline Implementation using components like Python Flask for API, Docker, CI/CD and AWS GPU Instance.
3. Worked on reducing response time of model by validating different system like Ram or GPU configurations. Response time reduced to 1-2 sec per paragraph from 2-3 minute by using GPU enabled instance.

Data Science Intern

ABI Health | From April 2019 to May 2019

BRCA Test Prediction (Oncology)

BRCA Test used for detection of Breast and Ovarian cancer. In this project data used is insurance claim data of patients. Using patient journey (i.e. claim data) predicting whether he or she requires undergone to BRCA Test. Stacked **GRU** Model architecture used for prediction of BRCA Test.

Malignant Nodules Classification and Detection from X-Ray

Classifying the malignant nodules in Chest X-ray images as abnormal using state of the art **CNNs** like **DenseNet-121, ResNext50**, Transfer-Learning by **fast.ai** libraries to help Radiologists (**89% AUC** is achieved in classification which is the best in industry).

Data Scientist

Shyena Tech Yarns | From Sep 2018 to March 2019

IOT Analytics

In this project created Dashboards for real time monitoring of Sensor Data. Collected data from sensors and push the data onto cloud. Analysed and Cleaning of acquire data was done. Identified various conditions from data using clustering algorithm.

Inventory Demand Forecasting

This product helps customers to maintain right inventory level of the products in the stock. The input is a Time Series Data related to the product sales and quantity sold. The product does data analysis, shows trends, seasonality. It also makes the data non stationary if required. The product supports various time series models for accurate forecasting (e.g. **S-ARIMA, Holt-Winters, LSTM, and XGBOOST**). The user shall have an option to apply a model of choice, compare the forecasting accuracy and then finalize the model.

PERSONAL PROJECTS

Rakuten Supervised Recipe Food Image Classification Competition

This was the competition organised by Rakuten. Task here was to identify various food items from image recipe. **CNN with transfer learning** was used for feature extraction and identification of food items. Our team secured **59th rank** in the competition.

Bottle Counting in manufacturing Industry using computer vision approach

Developed a computer vision model for counting the number of bottles manufactured every day in the industries. **Faster RCNN** was used for the model detection. A counting algorithm was implemented which gave us the actual count of the bottles.

Amazon Fine Food Reviews

Amazon Fine Food Reviews is a classic **Sentiment Analysis** problem used to classify the polarity of the review given by the Amazon user. Given the textual reviews and related features of the product, I have designed various techniques to classify the polarity of the review.

Quora Question pair Similarity

The Quora question pair similarity is to identify question asked on Quora by user is similar to previously asked question or not. In this Text processing and machine learning algorithm like **SVM, Logistic Regression and XGBOOS**T are used to classify questions are similar or not.

EDUCATION

M.Tech in Electronics (2016-2018) From VJTI Mumbai with CGPA 8.52.
B.E in Electronics and Telecommunication From SKNCOE, Pune with 63.61%

Personal Details

Date of Birth: 01/05/1993
Languages Known: English, Hindi, Marathi
Permanent Address: Plot no 92, Gut no 92, Oberoi Nagar, Satara Parisar, Aurangabad, Maharashtra.