Namratesh Shrivastav

Data Scientist

Email Id: namratesh@outlook.com
Mobile: +91- 9981867769
LinkedIn: linkedin.com/in/namratesh/
GitHub: github.com/namratesh
Medium: namratesh.medium.com/ (200K+views)

Professional Summary

Experienced IT professional with approximately 6+ years of expertise in machine learning. Proficient in leveraging Natural Language Processing (NLP) and deep learning methodologies, backed by a solid foundation in mathematics. Actively seeking opportunities to apply advanced NLP techniques for extracting and analyzing data from diverse sources, including PDFs, images, and text files, within a challenging and rewarding role.

Skills

- **Programming Language**: Python, SQL
- **Framework:** Tensorflow, Keras, Selenium, Pytorch, Streamlit, Gradio
- **Libraries:** Pandas, Scikit-learn, Seaborn, Ray, Celery, Pymongo, PyMuPdf, PaddleOCR, Tesseract
- **Modeling Skills**: Regression, Classification, Computer Vision, Natural Language Processing(including Langchain, Generative AI)
- **Cloud:**
  - AWS - Sagemaker, Lambda, S3, Cloud Watch, API Gateway, ECR, Endpoints Creation, Textract, Comprehend, EC2
  - AZURE - Blob, Kubernetes services , Logic Apps, Azure AI Foundry,
  - GCP - Vertex AI , GCR
- **DevOps/MLOps:** Docker, MLFlow, Git, DVC, RabbitMQ, Redis, Jenkins, Prometheus, Grafana

Professional Experiences

Publicis Sapient

July - Present

RAG Solution with Azure OpenAI for Medical Policy Data

- Built and deployed a healthcare RAG system on Azure using OpenAI and LangChain, which successfully **reduced support ticket volume** and **improved agent response times.**
- Implemented an advanced evaluation method (SAFE framework) that cut the launch time for new regions by 75% (**from 4 weeks to 1 week)**, saving thousands of developer and testing hours.
- Created an automated **evaluation pipeline on Kubernetes t**hat ran daily tests on over 300 questions to ensure the system's accuracy and reliability.
- Set up a CI/CD workflow with GitHub Actions t**o automatically test system changes**, making the validation process faster and more efficient for the development team.

Infrablok

Apr 2023 - June 2024

Detection of Sensitive Information and Masking while preserving the format

- Spearheaded a cross-functional team as the product owner, overseeing AI, DevOps, frontend, backend, and testing disciplines.
- Trained Named Entity Recognition (NER) models using Spacy, applied regex for pattern identification, and utilized dependency parsing to **detect PHI, PII, GDPR, and PCI** entities across diverse data formats such as **text, images** using **PaddleOCR, PDFs** using **PyMuPDF, and documents** using **Python-docx.**
- Achieved notable F1 scores of **81%, 87%, 82%, and 91%** for PHI, PII, GDPR, and PCI detection respectively.
- Implemented scalable solutions utilizing **Celery**, **RabbitMQ**, and **Redis** to process over 100GB of data within a couple of hours efficiently.
- Deployed data masking techniques to obfuscate sensitive data while **preserving the original format of PDFs, images, and documents.**
- Developed a **FastAPI** to identify sensitivity levels within folders, extracting data from MongoDB, and presenting findings on the user interface.
- Leveraged **Rclone** as a connector application for efficient data processing from S3, Google Drive, Azure Blob, and GCP storage solutions.
- Expanded the project scope by developing a **PromptGPT** system, employing **langchain and Mistral AI** increasing the overall detection and managing large requests. This system masks sensitive data before transmission to ChatGPT and restores it upon receiving responses.
- Employed **MLFlow** for model version control and management within the machine learning pipeline.

- Orchestrated the deployment of **multi-container Docker** configurations for each component, ensuring seamless automatic scaling.

TheSmartCube

Nov 2021 - Apr 2023

Extraction of Clause and allowed or Prohibited commodities from PDF

- Developed a solution to **remove strike-through tex**t from images using OpenCV.
- Used Textract API to **extract text from images** and store it in a text file.
- Built an AWS **Lambda function to extract exclusion clauses** from raw text using Regex.
- Trained an XGBoost model to identify correct **exclusion clauses**, and **hosted it on an AWS endpoint** that can be called from the AWS Lambda function.
- Trained **a custom-named entity recognition (NER**) model using Spacy to predict commodities.
- Trained a **masked language model (MLM)** to create tokenization for a deep learning classification model
- Trained **Roberta classification model** using MLM tokenization to predict allowed and not allowed commodities from a text.
- Deployed the custom-named entity recognition (NER) model and the Roberta classification model (based on MLM tokenization) on **Amazon Elastic Container Service (ECS)** using Docker.

Financial Table extraction from PDFs for English and German Language

- Extracted and analyzed financial data from Annual Reports of US and German companies, including Balance Sheets, Cash Flows, and Income Statements.
- Utilized NLP techniques such as stemming, lemmatization, and data cleaning to pre-process data for analysis.
- Trained and implemented the **XGBoost model** to automatically identify page numbers for table extraction.
- Utilized **Tabula to extract** and organize data from tables.
- Implemented **Transformer for machine translation** of German to English for accurate data analysis.
- Developed and deployed a **comprehensive API using Django** to connect from the front end.

Accenture

July 2019 - Nov 2021

Smart Search Engine using NLP

- Successfully developed a model that enables efficient search within files with minimal latency.
- Built a comprehensive data pipeline capable of handling various file types (pdf, doc, ppt, txt, etc.) and extracting their contents, performing basic NLP preprocessing such as text cleaning, tokenization, stemming, and lemmatization.
- Implemented the calculation of tf-idf scores on files to evaluate the similarity of the given query using cosine similarity.
- Displays the files on the web with the most relevant/similar words to the query at the top of the search results.

Achievements :

- Achieved **2nd place in the Machine Learning Hackathon** at The Smart Cube.
- Secured a **top 10 ranking (top 0.003%)** in the Global Technology Innovation Contest 2020 competition

Education

- **Bachelor of Engineering** in Electronics and Telecommunication from MITS, Gwalior 2019

Certification

- Tensorflow Developer by deeplearnig.ai, Applied Ai Course, AWS Certified Machine Learning Specialty