Avishek Das

Linkedin: *linkedin.com/in/avi539*
GitHub: avishekdas539

Email : *avishekdas539@gmail.com*
Mobile : +91-833-785-6650
LeetCode : avishekdas539

EDUCATION

Indian Institute of Engineering Science and Technology, Shibpur — Howrah, Westbengal
Bachelor of Technology In Civil Engineering ; CGPA: 9.15/10.0 — *2018 - 2022*
Key Courses Taken: Engineering Mathematics, Uncertainty Quantification, Structural, and Geotechnical Engineering

PROFESSIONAL EXPERIENCE

Tata Consultancy Services — Kolkata, India
Data Scientist — *Aug 2022 - Currently Working*
System Engineer

- Developed an advanced **GenAI-powered CV Screening System** with a focus on single responsibility, effectively segmenting tasks through **prompt engineering**. Implemented robust **data validation and sanity checks** using **Python** and **regex**, ensuring the integrity of LLM generated results. Conducted thorough evaluations of LLMs like **LlaMa-2-70B, IBM-Granite, GPT-3.5**, and **Mixtral-7B**, refining output precision. Utilized Azure AIML and IBM WatsonX platforms for development, resulting in a significant **accuracy increase from 60% to 85%** compared to the previous solution. Developed **Flask**-based **RESTful API** to serve the CV Screening Core Pipeline to integrate seamlessly with the fronted CORS policy.
- Developed an automatic sandbox for processing unorganized data and uploading to SQL server using Python. Developed the core part for formatting the unorganized data and convert it to organized and processable data frames using pandas. Applied different string manipulation techniques and **regex** to improve the data quality and validation. Optimized the sandbox core and was able to **reduce the processing time by one-sixth** which has removed the blockers at this stage for the client.
- Analyzed and performed extractive text summarization of financial reports using token **dependency mapping, Named Entity Recognition** and Natural Language Processing. Prepared word cloud to visualize the important words depending on frequency.
- Utilized **Natural Language Processing** techniques within the HSE sector to predict future incidents by analyzing historical data. Conducted topic modeling on free-text data to categorize incident causes and employed **sentence-transformer** to assess similarities between incidents for prediction. Developed a Flask-based web application integrating these models and statistical insights from historical data to proactively inform users about potential upcoming incidents.

PROJECTS

- **TinyBPE - Trainable Tokenizer based on Byte-Pair Encoding similar to GPT-2 and GPT-4** — June 2024
  - Implemented **trainable Byte Pair Encoding algorithm** based tokenizer from scratch using Python considering both byte level splitting and REGEX based splitting.
  - Implemented a multilingual tokenization issue for REGEX-based splitting to improve encoding and training quality.
  - Improved serialization and deserialization methods by eliminating dependency constrain.

- **STAAD-Ninja a Gemini-Pro Powered RAG Based Chat Bot for STAAD.Pro Solutions** — January 2024
  - Extracted and pre-processed texts from 7000+ pages of STAAD documentation and guide books. Performed chunking and semantic vectorization using **embedding-001** model by Google and stored into **FAISS vector database** for **RAG** using **LangcChain** framework. Leveraged **prompt engineering** techniques to optimize output from **Gemini-Pro** model. Integrated **multi-modal GenAI** in the app using **Gemini-Pro-Vision** model with the capability to remember the previous chats by implememting **custom Chat History module**. Deployed with **streamlit** based UI on streamlit apps.

- **American Sign Language Detection - ConvLSTM Based Solution** — November 2023
  - Developed a **1D Convolution and LSTM-based encoder-decoder model** to detect American Sign Language gestures by tracking hand landmarks, establishing a robust baseline for ASL detection systems. Implemented **NLP** tasks such as **tokenizing and embedding** for the decoder, achieving a **training accuracy of 98%** and **testing accuracy of 84%**. Contributed to the advancement of ASL recognition and accessibility technology. Demonstrated expertise in **signal processing** and NLP within the context of ASL gesture detection.

- **Real time Driver Alertness Monitoring Using Computer Vision** — August 2023
  - Prepossessed the image-labeled data for the data pipeline using TensorFlow Dataset API to train the machine learning models. Used custom layers on top of pre-trained MobileNetV2 model for faster classification using depth and point-wise convolution. Trained the custom layers to achieve **99.4% training accuracy & 96.8% validation accuracy.**

- **Emotion Detection from Texts using Deep Learning**   May 2023
  - Prepossessed textual data by tokenizing, removing stop words, removing numbers, expanding short contractions and lemmatizing. Also converted text to vector using Tf-IDF and word embeddings.
  - Created classification model using DecisionTree, RandomForest algorithm with 82% and 84.1% accuracy. Trained Embedding and LSTM-based deep learning model with a training accuracy of 92.9% and validation accuracy of 89.6%.

- **Automation in Structural Engineering**   Jun 2022 - Ongoing
  - Developed a deep learning-based regression model inspired by wave-net to predict the response of buildings due to wind and gravity load.
  - Developed model to predict optimum structural section sizes for a given condition using the KNN-Regression technique with 99.5% success rate of being the optimum section.

SKILLS SUMMARY

- **Programming Languages**: Python, Java, JavaScript
- **Machine Learning/ Deep Learning/ GenAI:**: Generative AI, LLM, Retrieval Augmented Generation, OpenAI API, Regression Models, Support Vector Machine (SVM), K Nearest Neighbours Clustering, Artificial Neural Network (ANN) ,Convolutional Neural Networks, RNN/ LSTM/ GRUs, BERT/ Transformers, Natural Language Processing, Computer Vision, Deep Neural Network, Clustering Based Algorithm, Generative AI/ LLM
- **Frameworks/ Libraries**: TensorFlow, PyTorch, nltk, spacy, scikit-learn, LangChain, FAISS VectorDB, Gensim, BeautifulSoup4, numpy, pandas, pyspark, scipy, matplotlib, plotly, seaborn, flask, streamlit, SQLite3, HTML, CSS
- **Tools/ Platforms**: Dataiku (Core Designer Certified), Advanced Microsoft Excel, Microsoft Word, Microsoft PowerPoint, Azure AIML, IBM WatsonX.AI
- **Version Control**: Git, GitHub

INTERNSHIP

- **Summer Research Intern - IIEST, Shibpur**   May 2020 - July 2020
  - Worked on vehicle detection and tracking using YOLO pre-trained model to extract traffic-related data and movement curve of traffic on some given road.

AWARDS

- **WatsonX Assistant Challenge for GSI - IBM - 4th Position**   2023
  - Developed a chat bot using WatsonX Assistant on IBM Cloud platform by defining custom actions and their steps for banking specific industries.

- **National Award Competition (Civil) - Institute for Steel Development & Growth - 1st Position**   2021
  - Designed "Iconic Steel Roof Structure over An Open-Air Theatre and Stage" with a team of 4 people. Awarded the first position at the national level

- **Bentley Future Infrastructure Star Challenge - Bentley Systems - Selected in top 20 projects**   2021
  - Innovative Idea Presentation - Energy extraction from the turbulence of wind due to high-speed vehicle. Worked in a team of 2 people. Selected In Top 20 Projects.

- **Build O' Innovate (Smart Traffic) - Indian Institute of Technology, Patna - 3rd Position**   2021
  - Development of an automatic flow meter to measure traffic parameters using YOLO pre-trained models. Participated with a team of 3. Awarded second place at the national level.

COURSES/CERTIFICATIONS

Dataiku - Core Designer. | Natural Language Processing with Python. | Neural Networks and Deep Learning. | Improving Deep Neural Networks: Hyper-parameter Tuning, Regularization and Optimization. | Data Structures & Algorithms - Java.