ASHUTOSH SINGH

Erlagen, Germany.
☎ +49 17670784103| ✉ ashutosh.singh.de@gmail.com | Github | StackOverflow

Professional Profile:

I am currently pursuing my Masters in Computational Engineering from Friedrich Alexander University. I have 3+ years of experience working with AI-Startups working in the field of conversational AI.

Skills Summary:

- Statistics
- Machine Learning
- Data Visualization
- Automation
- Data Engineering
- Text Ranking
- OpenCV
- Neural Networks
- Django
- django-rest-framework

- Python
- Containerized Applications.
- Serverless Architectures
- NLP
- Text Analytics
- Elastic Search
- Keras
- PyTorch
- Bayesian Learning

Relevant Studies:

FRIEDRICH ALEXANDER UNIVERSITY, ERLANGEN AND UNIVERSITÀ DELLA SVIZZERA ITALIANA, LUGANO

Deep Learning
- We implemented networks like FNN, CNN and RNN from scratch in python.
- Basics of object detection and segmentation.
- Architectures like ResNet, LeNet, U-net.
- Implementations of various layers and Units

Advanced Topics in Machine Learning
- Overview of GANs and VAEs
- Text mining and Natural Language Processing
- Overview of various topics from the domain of Deep Reinforcement Learning.
- Reproducibility report on paper "Early Stopping in Deep Networks: Double Descent and How to Eliminate it"

Machine Learning in Signal Processing
- Algorithms like Linear, Logistic regression and their dissection
- Basics of convex optimization problems
- Wavelet Transform and basics of compressed sensing.
- Basics of Neural networks and SVMs

Image and Video Compression
- Algorithms like Linear, Logistic regression and their dissection
- Basics of convex optimization problems

High Performance Computing
- Working with OpenMP and OpenMPI APIs to implement parallel programs
- Python based pyMPI to implement various parallel programs.

Computer Vision
- Advanced computer vision course
- Currently taking this subject.

Linear Algebra
- Advanced Linear Algebra course at FAU
- Topics like EigenValue problems, Krylov Space methods

Intro To Bayesian Computing
- Introduction to sampling: Importance Sampling, Rejection Sampling, Gibbs Sampler etc
- Monte Carlo and Markov-Chain-Monte-Carlo
- [Project Report](#) on User Clustering based on website usage.

Experience:

FRAUNHOFER INSTITUTE OF INSTITUTE FOR INTEGRATED CIRCUITS - RESEARCH ASSISTANT (GERMANY, ERLANGEN|NOV 2021 – CURRENT)

Efficient Convolutional Networks
- Studying and analyzing latest developments in efficient architectures of Image Segmentation.
- Ongoing Project: Create UNet like architectures based on more efficient and sparse architectures like Depth-Wise separable and grouped convolutions.

JIO HAPTIK TECHNOLOGIES LIMITED(HAPTIK) - MACHINE LEARNING ENGINEER (INDIA, MUMBAI|JAN 2019 – JAN 2021)

Domain Classification System
- Domain Classification/Tagging based on entity embeddings.
- Parsed 400GB freebase data to extract entities and their definitions/descriptions.
- Finetuned GloVe vectors to include these entities.
- Trained a text classification system using Flair and pooled embeddings to pool BERT outputs with custom GloVe vectors

Short-text Similarity
- **LSTM + Neural Tensor Network**
  - Model to interpret semantic similarity between two short excerpts of text(messages in a chat)
  - It consisted of an LSTM encoder followed by a [bilinear-layer(tensor layer).](#)

- It was trained specifically on chat data.
- Pretrained glove(twitter) embeddings were first fine-tuned on 200MB of textual chat-like data For eg. youtube comments, twitter conversations, daily-dialog, etc. and then used while training the model.
- The model was able to differentiate between adversarial sentences like "I love you" and "I hate you"; this is one of the cases where existing sentence encoders fail.
- This architecture was used to deploy models for Smalltalk/chit chat for 10 languages.

Paraphrase Generation
- **Variational Autoencoders**
  - Variational autoencoder was used to build an architecture proposed by Gupta et al, A Deep Generative Framework for Paraphrase Generation.
  - It was trained using quora similar questions dataset.
  - Github link - https://github.com/ashutoshsingh0223/deep-paraphrase

Sequence Tagging and Named Entity Recognition
- Designed and implemented Custom Embedding dropout for feature based CRF models
- Multiple OOV vector assignment strategies.
- Short text NER
  - This architecture was inspired by a Spotify report.
  - They were using stochastic methods which only considered prefixes and suffixes of an entity for detection
  - The entity itself was completely masked while training.
  - This method proved quite useful when handling sequences of short lengths.
  - Github link - https://github.com/ashutoshsingh0223/short_text_ner
- BERT-CRF
  - Tested with the SNIPS Custom-Intents dataset.
  - 96% Accuracy.

Oracle
- A step towards efficient data centric model development
- Diagnostics module for intent detection system at Haptik.
- Data for false positives, outliers for each intent and intent/class overlap for the underlying text classifier
- Deliver this data to bot-developers through SES.
- Oracle runs as a Celery task
- Oracle was built as a step towards fitting data to a general classification algorithm, as opposed to Common tasks framework where you try to find the best algo for a dataset.
- Oracle scans your training data to give you hints on how to improve your data so the haptik's intent classification can more easily understand it

JUBI.AI - DATA SCIENTIST (INDIA, MUMBAI|JUNE 2017 – DECEMBER 2018)

DIY Bot Building Platform
- Built in-house intent recognition system.An ensemble of text ranking and text classification algorithms. The bot builder has been used to build scalable chatbot solutions.

- Able to handle more than 300 intents, bot builder is trusted by serious banks and mutual fund companies to strengthen their customer service. Transactions have been made super easy with these bots.
- Deployed the machine learning module as a combination of five microservices on a SERVERLESS architecture.
- Microservices were containerized using Docker and deployed on AWS.

Customer Chat Analytics
- Jubi.Ai helps various banks, insurance companies, and mutual fund companies to identify the redundant/repeated topics in their customer service queries and automating most of their customer service thus save a lot of money for them.
- My role was to provide them with insights from the chat logs between agents and users by using various topic modeling and information extraction techniques.
- Used various phrase extraction techniques like chunking and noun-phrase extraction.

Data Collection
- Scraping websites for text related to mutual funds, personal finance, life insurance, and banking.
- Creating clusters of queries from various platforms to get an insight into user behavior.
- Training word embeddings like word2vec to boost the learning capacity of bots.
- Producing roadmap/view to creating an exhaustive set of pre-trained embeddings and language models around text related to the financial domain

OpenCV Image Processing
- Built a tool to get structured data out medical pdf reports.
- Pdfs are encoded documents which when converted to text lose their structure.
- We converted it into an image and tried to fit a tabular structure around medical test names, values, units and normal range of that value.
- This fitting was greedy in itself and found the best frame-fit by minimizing a custom error function after manually tweaking some hyperparameters.

Real-Time Log Analytics and Bot-retraining System
- Used MongoTriggers to stream data in real-time from live mongo database, transform it in a column-row DB compatible schema and store it in a MongoDB cluster.
- Used dash(a framework to build analytical frontends in python) and flask with pandas to built a reactive frontend for the retraining pipeline.

Site-Search for HDFC Mutual Funds
- Extracted the website data scattered across multiple SQL tables.
- Indexed it in a bm25 text search engine.
- Hosted the search engine with ECS.

Education:

- **Bachelors in Technology (BTech)** – Electronics and communications
  KNIT,SULTANPUR, UP, India **(81%)**
- **Senior School Certificate Examination(CBSE) - 92.6%**
- **Master of Science - Computational Engineering, Information Technology - Digital Signal Processing** Friedrich Alexander Universitat, Erlangen, Nuremberg, Germany. 2020-2022
- **Master of Science - Computational Science** USI(Università della Svizzera Italiana), Lugano, Switzerland. 2020-2022

Relevant Courses:

- Kubernetes the hard way
- Google cloud certified data engineer
- Elastic Stack Essentials
- SQL

Extra Accolades:

- Scored **99 percentile** in **GATE**(National Level Exam for Engineering Graduates in India)

- Got a chance to study at IIT Madras(Control Systems) and IIT BHU(Signal Processing)