REYAN ZAFIR

(+91) 9800456222
reyanzafir@gmail.com

Data Scientist

LinkedIn: reyanzafir

Data Scientist with advanced technical proficiency in building data-driven applications and has considerable expertise in formulating data science solutions in diverse industries and identifying cost-saving opportunities for businesses. Experienced in Predictive Modelling, Data Preparation and driving sustainable adoption of ML applications using best practices. Working predominantly in Machine Learning, Deep Learning and Natural Language Processing(NLP) space.

Education

Indian Institute of Technology, Kharagpur, *Integrated B.Tech+M.Tech (Dual Degree)* 2015 — 2020

Work Experience

Optum, United Health Group | Data Scientist **August 2020 — Present**
DPaaS: Data Protection as a Service *Bengaluru*

- Built a deep column analyzer that flags out **PHI/PII** columns **(HIPPA compliance)** in tables residing in data lakes to reduce **high TAT** and human errors by developing **in-house supervised DL based algorithms** tuned for yielding high recall score
- Reduced estimated time to complete by **11 folds** by assisting data stewards in their manual attestation process
- Designed a **many-to-one sequence model** that essentially learns the inherent and positional feature of raw values in tables
- Built a **CNN based architecture** while adding global features and **SymSpell algorithm** to have a generic representation
- Containerized the solution to be used through **GIT webhook** for consuming tables and hosted DL models on **TFServing**

Context-Driven Spell Correction Module for OCR engine

- Designed a self-supervised context driven spell correction module by fine-tuning **BERT based masked language model**
- **Reduced WER by 60%** while spending 1% extra time in spell correction, 60 times faster than a conventional spell checker
- Performed **Transfer Learning** on 1k domain-specific documents that resulted in reducing perplexity from 12.79 to 2.35
- Built an **image-orientation correction module** to enhance OCR performance and achieved an accuracy of 98.7%

Internships & Projects

Optum, United Health Group | Data Science Intern **May 2019 — July 2019**
Automatic Speech Recognition for monophonic voicemail dataset *Bengaluru*

- Designed a **DeepSpeech2** version for call records with missing frequencies domain knowledge and achieved a CER of 33%
- Affixed **SpecAugmentation** to handle the deformation across time and frequency information of audio clips
- Coalesced band-pass filter, augmentation policy, amplitude amplification to enhance robustness for different domain audio
- Modelled **JASPER**, a deep time-delay neural network and performed Transfer Learning which resulted in a **CER of 17%**

SymphonyAI | Data Science Intern **May 2018 — July 2018**
Embedding space for medical domain entities *Bengaluru*

- Implemented **word2vec, GloVe** and various other embedding techniques for miscellaneous downstream tasks in NLU
- Initiated Knowledge Graph construction by incorporating coreference resolution in the relationship extraction for entities
- Developed relationship embeddings using Knowledge Graph for feature representation by establishing the entities' relation

Bachelor's Thesis Project | Prof. Jiaul H. Paik, IIT Kharagpur **August 2018 — March 2019**
Diagnostic Analysis of Neural Machine Translation Models

- Designed an encoded-decoder type architecture for sequence generation and analyzed the shortcomings for long sequences
- Developed an **Attention-Based** Neural Machine Translation model for bilingual translation and achieved an **accuracy of 88%**
- Evaluated the significance of **attention-weights** and the performance of a bidirectional LSTM network for machine translation

Term Projects & Hackathons | IIT Kharagpur

- Modeled a **movie recommendation system** and achieved an accuracy of 0.890 using K-Means and Matrix Factorization
- Built a **CHURN prediction** model by using Random Forest algorithm with an accuracy of 92.6% for a telecom company
- Designed a **Credit Card Default** prediction model using XGBoost and secured 34th rank among 600+ participants

Skills

Programming Languages	Python, R, Cypher, GSQL, C, C++
Software/ Tools	FlaskAPI, Docker, Kubernetes, Neo4j, TigerGraph, GIT
DL Frameworks | Libraries	TensorFlow, Keras, PyTorch | Pandas, sklearn, NumPy, OpenCV, Gensim, NLTK, LibROSA

AWARDS & ACHIEVEMENTS

- Awarded Star Performer of Q1' 2021 at Optum for a commendable job in owning key components of deliverables
- Incubated under HSBC for proposing an AI-based solution for a social cause – December 2018
- Received best All-Rounder fresher for meritorious services to the Radhakrishnan Hall of Residence during the session 2016-17