+91 98XXX XXXXX
contact@subhalingamd.me
Chennai, IN

Subhalingam D

subhalingamd.me
github/subhalingamd
linkedin/subhalingamd

Education

- **Indian Institute of Technology, Delhi** **Jul 2018 – May 2022**
  Bachelor of Technology in Mathematics and Computing | *CGPA: 8.196/10*
- **Chennai Public School, Chennai** **Jun 2016 – May 2018**
  All India Senior School Certificate Examination (CBSE Std. XII) | *Percentage: 96.4%*
- **Chennai Public School, Chennai** **Jun 2014 – May 2016**
  All India Secondary School Examination (CBSE Std. X) | *CGPA: 10/10*

Experience

KnowDis Data Science, Delhi **May 2022 – Present**
Data Scientist

Product Category Search Engine (for IndiaMART)

- Observed recall@2 of 94% (**+6% than recall@1**) and motivated to build a system to rescore top-*k* categories for improving accuracy
- Built a reranker that **encodes the query & retrieved categories independently**, aligns each query token with the most relevant category token and aggregates the similarity scores across the query; the category embeddings are **pre-computed** offline and cached in memory
- Revised confidence classification rules, resulting in 82% (**+6%**) **coverage** for **high-confidence** class while maintaining accuracy at 95.5%
- Attained a 1-2% gain in overall accuracy and currently working on parallelizing the encoding step in the reranker with the retriever

Query Understanding (for IndiaMART)

- Developed a two-stage system to **identify all the relevant attributes** mentioned in a query and **extract their corresponding values**
- Trained **BART** & **RoBERTa** models using product specifications data with auxiliary category classification task to give more supervision
- Formulated a negative sampling strategy and **incorporated additional embeddings** to tackle **incomplete tagging** in the training data
- Deployed the system using **FastAPI** and presented a **demo to the client**; planned to integrate with search system for **refining results**

English-to-Hindi Translator with Style Restriction

- Built an **mBART**-based translation baseline for converting **English texts to Hindi** in a **specified style** using in-house parallel corpora
- Obtained the English translations for scraped monolingual Hindi data using **Google Translate API** to augment the training data
- Reviewing existing works on controlling styles in text generation, specifically for low-resource settings, to create improved systems

Other contributions:

- Explored non-autoregressive generation methods to convert Roman Hindi words in search queries to English to achieve low-latency
- Experimented with lexical string matching using Elasticsearch to handle model numbers in a search query

KnowDis Data Science, Delhi **Jan 2022 – May 2022**
Data Science Intern | *Product Category Search Engine (for IndiaMART)*

- Devised an NLP scheme to predict the most relevant **product category** (from **113k possible labels**) from user queries/product listings
- Trained a transformer-based **classifier** on automatically labelled data and added heuristics to improve knowledge of category labels
- Incorporated **causal attention mask**, which improved results; fine-tuned **T5** model for **oversampling** under-represented categories
- Achieved similar accuracy (**~88%**) as the previous seq2seq model while significantly **reducing average response time** (**3x faster**) and completely **eliminating timeouts**; the model was **integrated with *IndiaMART*'s search system** and was deployed in production

Samsung R&D Institute, Delhi **Jun 2021 – Jul 2021**
Software Engineering Intern | *Acoustic Sound Source Localization, Tracking and Separation*

- Developed **sound source direction estimation** module using time delay of arrival of signals between pairs of microphones in an array
- Added modules for tracking active sound sources and extracting individual signals for downstream object identification pipeline
- Integrated stationary noise estimation module for ambient noise removal and reduced maximum direction of arrival error to 7°
- *Received **Pre-Placement Offer (PPO)** for impeccable performance during the internship*

MateRate Education, Delhi **May 2020 – Jul 2020**
Machine Learning and Web Development Intern | *Students' Latent Knowledge Space Modelling and Results Portal Development*

- Developed Item Response Theory-based models to estimate and analyze the **ability** of 5000+ students & **difficulty** of 200+ questions
- Designed database schema and built Web APIs using **Django REST framework** to display students' performance reports to parents
- Deployed *Django* backend using **Elastic Beanstalk** with *MySQL* on **RDS** and *React* frontend to **S3** with **CloudFront** CDN integration
- Set up **Auto Scaling group** and attached **Load Balancer** for horizontal scaling; the portal went live with the results of 5000+ students

Skills

Languages: **Python**, **Java**, C++, C, Bash, MATLAB
Deep Learning: **PyTorch**, Transformers, PyTorch-Lightning, Accelerate, TensorFlow, Keras, NLTK, spaCy
Development: **FastAPI**, Django, **AWS**, Streamlit, SQL, CSS, jQuery, HPC Cluster, Docker, Git

+91 98XXX XXXXX
contact@subhalingamd.me
Chennai, IN

Subhalingam D

subhalingamd.me
github/subhalingamd
linkedin/subhalingamd

Projects

Tracking State Changes for Entities in Technical Procedural Text **Feb 2021 – Apr 2022**
Prof. Srikanta Bedathur and Prof. Maya Ramanath, Research Project (under IBM AI Horizons Network) *[Paper]*

- Prepared a dataset consisting of ***How-to*** troubleshooting FAQs by **scraping WikiHow pages** from *Computers and Electronics* category
- Constructed **BERT**-based baselines to **predict changes in properties of the entities** involved at each step of the process
- Surveyed the literature to build **next-step recommender** from a given sequence of performed actions and developed LSTM baselines

Identification of Hate Spreaders on Social Media **Jan 2022 – Apr 2022**
Prof. Niladri Chatterjee, Bachelor's Thesis *[Thesis]*

- Identified key features to profile hate spreaders on Twitter using their feeds and observed high feature importance for sentiment scores
- Proposed a novel scheme that uses **GloVe** embeddings for encoding and **sentiment scores** as weights to mark word importances
- Attained an **accuracy of 76%** (for English language) on the *PAN@CLEF 2021* dataset *(+1% than best)* and 77% with an ensemble

Multilingual Question Answering **Oct 2021 – Nov 2021**
Prof. Mausam, Natural Language Processing Course

- Utilized **XLM-RoBERTa** model for question-answering in **Hindi & Tamil** to **predict the answer span** in a context for a given question
- Fine-tuned on *chaii-1* + *MLQA* + *XQuAD* (for Hindi) + Google translated *SQuAD* (for Tamil) datasets; attained Jaccard score of 68.72%

Rule-based Written-to-Spoken Text Converter **Aug 2021 – Sep 2021**
Prof. Mausam, Natural Language Processing Course

- Developed a system to rewrite sentences in the spoken form, accounting for dates, times, abbreviations, numerical quantities & units
- Refined the rules through iterative error analysis to handle edge-cases (like inflections) and achieved **97.94% F1-score (top in class)**

Corporate Bankruptcy Prediction **Feb 2021 – Apr 2021**
Prof. Niladri Chatterjee, Data Mining Course *[Report]*

- Inspected bankruptcy prediction models and observed poor recall; hypothesized **class imbalance** & **missing values** to be the reasons
- Trained an ensemble model with **Mean Imputation** & **SMOTE** transformations on *Polish companies* dataset and **gained +10% recall**

Extended Vector Space Model for News Articles Retrieval **Oct 2020 – Nov 2020**
Prof. Srikanta Bedathur, Information Retrieval Course

- Created an end-to-end retrieval system indexed using **TF-IDF weights** with support for prefix search & named-entity based filters
- Reduced index size by half with **gap encoding**; applied pseudo-relevance feedback based probabilistic **query expansion** for reranking

More projects:

- **Context-Sensitive Word Sense Disambiguation**: Studied disambiguation capability of *BERT* and *GloVe+BiLSTM* using *WiC* dataset
- **Tweet Sentiment Classifier**: Vectorized tweets using *TF-IDF* after pre-processing and fed into an *LR* classifier; attained 78.33% accuracy
- **Movie Recommender System**: Implemented collaborative filtering model using SVD on MovieLens dataset, resulting in MAE of 0.6671
- **Adaptive Neuro-Fuzzy Inference System for Diabetes Prediction**: Trained a *Takagi–Sugeno* type system with an accuracy of 81.3%

Relevant Coursework

Natural Language Processing, Information Retrieval and Web Search, Data Mining, Linguistics (*via* Intro to Language Sciences; Language and Communication), Data Structures and Algorithms, Probability and Stochastic Processes, Statistical Methods, Linear Algebra, Calculus, Fuzzy Sets and Applications, Operating Systems, Differential Equations, Optimization Methods, Theory of Computation

Activities

- **General Secretary,** Mathematics Society, IIT Delhi Aug 2021 – Jun 2022
  Led a team of 30+ members in organizing events, workshops, blogs & interviews to foster the development of maths among students
- **Teaching Assistant,** Information Retrieval and Web Search, *Graduate course offered by Prof. Srikanta Bedathur* Aug 2021 – Dec 2021
  Responsible for preparing rubrics for exams & assignments, grading them and assisting with projects
- **Web Developer,** Student Activity Council, IIT Delhi Feb 2020 – Sep 2020
  Revamped the website's design and functionality to enhance user experience and make information more easily accessible
- **Web Development Executive,** Entrepreneurship Development Cell, IIT Delhi Sep 2019 – Jun 2020
  Collaborated on the design and development of an Entrepreneurship Portal for the Institute
- **English Language Mentor,** Board for Student Welfare, IIT Delhi Aug 2019 – Dec 2019
  Helped newcomers improve their English language communication skills through regular interactive sessions
- **Volunteer in Teaching projects,** National Service Scheme (NSS), IIT Delhi
  Dedicated 120+ hours of community service by tutoring school students and creating educational videos for various subjects