Indraneil Paul

Applied Scientist @ Amazon | CS Dual-Degree @ IIIT Hyderabad

Email Github Scholar LinkedIn Website

Work Experience

April 2020 – Present — Applied Scientist , Amazon, Bangalore

- Part of the Content Moderation Science team at Amazon Performance Advertising
- Experimented with multi-task multi-lingual pre-training regimes for Transformer-based networks
- Deployed an automated moderation solution as part of Advertising expansion to European marketplaces which increased automation by 28% while intercepting forbidden ads at a 17% higher rate
- Working on sample-efficient approaches to train high-capacity Vision Transformer models using subtask distillation

HuggingFace Transformers | FastAI | PyTorch | Python | CUDA C++

August 2019 – March 2020 — Software Development Engineer , Amazon, Hyderabad

- Part of the Transportation Plan Coordinator team for container shipments at Amazon Logistics
- Designed and partially implemented a constraint-aware cost-optimal package drop-off planner for allowing merchants to rank options to schedule last-mile drop-offs
- Scaled an tier-1 asynchronous service for year-end peak including database tuning, JVM latency optimizations, profiler based host type changes and scaling queue workers

Spring | AWS SNS | AWS SQS | AWS DynamoDB | METIS | Java

June 2016 – August 2019 — Research Assistant , Language Technologies Research Center, IIIT Hyderabad

- Worked under the guidance of professor **Ponnurangam Kumaraguru** and **Manish Gupta**
- Characterized the connectivity and activity levels of verified users on social media platforms
- Employed temporal, network and content based features for a discriminative model
- Uncovered the aspects of a user's profile, topic usage and activity that best predict verification
- Released a fully featured **dataset** of 400k+ users, containing 79+ million edges and 494+ million Tweets

Graph-Tool | FastAI | Neo4j | NLTK | Twitter API | PoweRLaw | Python | R

July 2018 – July 2019 — Research Assistant , Machine Learning Lab, IIIT Hyderabad

- Worked under the guidance of professor **Sujit Gujar**
- Explored the use of two-sided matching algorithms on dynamic graphs
- Researched location constraint aware graph matching
- Formulated applications of constrained mechanism design in resource exchanges and ride sharing

ParamILS | CVXOpt | MATLAB | Python | C++

May 2017 – August 2017 — Google Summer of Code , Green Navigation, Netherlands

- Improved fuel consumption prediction in electric vehicles given intended route and associated terrain
- Added bayesian optimization for optimal model hyperparameter selection for legacy models
- Modified pre-existing machine learning solution to leverage LSTM based time series prediction and reduced absolute forecasting error by 39%

TensorFlow | Pandas | BayesOpt | Python

Education

July 2019 — **Masters by Research**, IIIT Hyderabad — CGPA 8.52

May 2017 — **Bachelors of Technology in Computer Science**, IIIT Hyderabad — CGPA 6.85

Summer Schools

September 2021 — Gaussian Process and Uncertainty Quantification Summer School (GPSS)

August 2021 — European Summer School in Logic, Language and Information (ESSLLI)

References

IIIT Hyderabad	**Ponnurangam Kumaraguru**, Thesis Advisor	Email
Amazon	**Sumit Negi**, Principal Science Manager	Email
Microsoft	**Manish Gupta**, Thesis Co-Advisor	Email

Publications

Data-Efficient Training of Vision Transformers via Distilling Prototypical Self-Labels
Amazon Machine Learning Conference 2021, Seattle
Indraneil Paul, Sumit Negi
Abstract | PDF

Multimodal Pose-Aware Pipelines for Sexually Explicit Advertisement Moderation
Amazon Machine Learning Conference 2021, Seattle
Indraneil Paul, Soumya Roy, Sumit Negi
Abstract | PDF

Multilingual Moderation Models for Sponsored Books Advertisements in Skewed Low-Resource Settings
Amazon Multilingual NLP Workshop 2021, Seattle
Indraneil Paul, Prasanna Krishnaswamy, Sumit Negi
Abstract | PDF

Insights into, Analyses and Prediction of Verified Users on Twitter
Masters Thesis, IIIT Hyderabad
Indraneil Paul
Abstract | PDF

What sets Verified Users apart? Insights, Analysis and Prediction of Verified Users on Twitter
WebSci 2019, Boston
Indraneil Paul, Abhinav Khattar, Shaan Chopra, Ponnurangam Kumaraguru, Manish Gupta
Abstract | PDF

Elites Tweet? Characterizing the Twitter Verified User Network
LSGDA Worskshop, ICDE 2019, Macau
Indraneil Paul, Abhinav Khattar, Shaan Chopra, Ponnurangam Kumaraguru, Manish Gupta
Abstract | PDF

Personal & Capstone Projects

Multilingual Language Detection — February 2021 - March 2021
Implemented a Collapsed Gibbs Sampling inference to model a multi-label language detector that can be trained on single-language documents. The model learns to approximate the generative process of texts in a specific language by modelling a document as a mixture of languages and a language as a mixture of words. The resultant model is able to detect mixture of languages in documents in an interpretable manner even when trained on small corpora.

HuggingFace Tokenizers | Python

News Article Autosummarization — January 2017 - May 2017
Implemented an Hierarchical LSTM based sequence to sequence model to automatically generate a grammatically coherent gist of a news article. Trained multiple sequence models in a topic aware manner aided by Hierarchical Dirichlet Process based multilevel topic models for sequence model selection.

Keras | Python

Kaggle Two-Sigma Financial Forecasting Challenge — January 2017 - March 2017
Implemented a pipeline to compare the efficacy of various time series forecasting approaches in accurately predicting the future value of various anonymized financial instruments. Utilized GARCH models and multiple boosting techniques, eventually settling on an ensemble.

TensorFlow | Statsmodels | XGBoost | ARCH | Python

Wikipedia Search Engine — August 2017 - November 2017
Developed a search engine using term-document techniques capable of ranked retrieval over body and header text, references, links and metadata on a Wikipedia XML dump. Enhanced the system to support caching and B+ Tree indexing to enable the system to scale to the entire Wikipedia 14 GB corpus.

BTrees | SAXParser | Python

Co-Authorship Link Completion — January 2016 - May 2016
Created a system that ingests a large number of Computer Science research papers from the DBLP archives and uses the set of topical tags, abstract and title corresponding to each paper, to predict a field in which a certain author is likely to contribute in the near-future. Extended the system to predict future academic collaborations based on topical fingerprints. Used the memory-mapped key-value store LMDB to facilitate sub-second retrieval.

TensforFlow | LibSVM | LMDB | Python