Gaurav Arora

APPLIED RESEARCH SCIENTIST @ AMAZON

9888552871 | contactgauravforwork@gmail.com | www.arggaurav.com | goru001 | gaurav-arora-23593220

Education

Georgia Institute of Technology

Atlanta, Georgia

MASTERS IN COMPUTER SCIENCE WITH SPECIALIZATION IN MACHINE LEARNING *2020 - 2022*

- My GPA was 3.8/4. I pursued my Masters remotely during Covid in addition to full-time job at Haptik and then at Amazon.

Punjab Engineering College

Chandigarh, India

BACHELOR OF TECHNOLOGY (B.TECH.), COMPUTER SCIENCE AND ENGINEERING *2014 - 2018*

- My CGPA was 7.77/10. I had the best JEE-Mains Rank in class.

Experience

Amazon

Bengaluru

APPLIED SCIENTIST *Apr 2021 - Present*

- Building and improving multiple NLU components (Intent, NER, Contextualization, Policy Engine) for Conversational Shopping Assistant. My work has resulted in improvements across coverage and precision of bot response upto 10%.
- Worked on improving representation of code-mixed data in Transformer models, paper for which was accepted in ACL Findings 2023.
- Also worked on forecasting problems using Deep Learning and built the demand forecasting system for predicting smartphones demand for Amazon India, improving the model performance by 20%

Haptik (part of Jio Platforms Limited)

Mumbai

MACHINE LEARNING SCIENTIST - 2 *July 2019 - Apr 2021*

- Built the Intent Detection System for Haptik's NLU Engine, owning it from Research to Production
- The performance of new system was bench-marked against 10 data-sets, including 3 open source datasets and it was 25% more accurate than their previous system, with latency under 150ms
- We also benchmarked the new system against popular NLU platforms like Dialogflow, RASA, LUIS in the paper HINT3: Raising the bar for Intent Detection in the Wild and Haptik's system achieved similar performance as others.

Goldman Sachs

Bengaluru

SOFTWARE ENGINEER *June 2018 - July 2019*

- Worked on Analytics for Desktop Assistant to bring to the fore key metrics about Product Health and User Behavior

Researchshala

Chandigarh

CO-FOUNDER *Nov. 2016 - Mar. 2018*

- Built an online platform to connect professors to research-interns, helping them with their research projects

Goldman Sachs

Bengaluru

INTERN *Jan. 2017 - June 2017*

- Worked with the team that built and maintains the firm's standard UI Development Framework - the UI Toolkit

Selected Publications

CoMix: Guide transformers to code-mix using POS structure and phonetics

ACCEPTED AT ACL FINDINGS 2023

- We propose CoMix, a pre-training approach to improve representation of code-mixed data in transformer models by incorporating phonetic signals, a modified attention mechanism, and weak supervision guided generation by parts-of-speech constraints.
- CoMix improves performance across four code-mixed tasks: machine translation, sequence classification, named entity recognition (NER), and abstractive summarization. It also achieves new SOTA performance for English-Hinglish translation and NER on LINCE Leaderboard and provides better generalization on out-o-fdomain translation.
- We also propose a new family of metrics based on phonetics and demonstrate that the phonetic variant of BLEU correlates better with human judgement than BLEU on code-mixed text

Natural Language Toolkit for Indic Languages - iNLTK

goru001/inltk

ACCEPTED AT EMNLP-2020'S NLP-OSS WORKSHOP *https://arxiv.org/abs/2009.12534*

- iNLTK provides out-of-the-box support for Data Augmentation, Textual Similarity, Sentence Embeddings, Word Embeddings, Tokenization and Text Generation in 13 Indic Languages
- iNLTK has 70,000+ downloads, 700+ stars on GitHub, and has been widely shared and appreciated on Twitter, LinkedIn, Reddit
- By using pre-trained models and data augmentation from iNLTK, we achieve more than 95% of the previous best performance by using less than 10% of the training data on publicly available classification datasets

HINT3: Raising the bar for Intent Detection in the Wild

hellohaptik/HINT3

ACCEPTED AT EMNLP-2020'S INSIGHTS WORKSHOP *https://arxiv.org/abs/2009.13833*

- This paper introduces 3 new datasets created from real user queries received by live chatbots, evaluates popular NLU platforms and highlights critical gaps in language understanding
- Paper introduces novel subset approach for evaluation and discusses trade-off between out-of-domain vs in-scope performance

Pre-training ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection

goru001/nlp-for-manglish,nlp-for-tanglish

ACCEPTED AT FIRE 2020 - FORUM FOR INFORMATION RETRIEVAL EVALUATION *https://arxiv.org/abs/2010.02094*

- This paper proposes pre-training ULMFiT on synthetically generated code-mixed data, generated by modelling code-mixed data generation as a Markov process
- Model achieved 0.88 weighted F1-score for code-mixed Tamil-English language in Sub-task B and got 2nd rank on the leader-board

Honors & Awards

Mar. 2021	**Indian Achievers Award 2020**, from Indian Achiever's Forum (IAF) in the Young Achievers Category for contribution in nation building through iNLTK	*India*
Mar. 2020	**Fast.ai International Fellow**, for contributions to Fast.ai forums	*Worldwide*
Mar. 2019	**Fast.ai International Fellow**, for contributions to Fast.ai forums	*Worldwide*
Dec.2018	**Top-17% , Human Protein Atlas Image Classification, Kaggle**, for developing Deep Learning model which classified mixed patterns of proteins in microscope images. The competition had 2172 teams, but I participated individually and hence had 100% contribution in the 366th placed solution	*Worldwide*
Oct.2017	**1st Prize,IEEE-Hackathon**, for developing chat-bot to help people with emotional decisions in life	*PEC*
Feb.2016	**Top-100**, among 500,000 students in IT-Olympiad,2016.	*India*
Oct.2016	**2nd-Prize, IEEE-Hackathon**, for developing an Augmented reality application to help teachers	*PEC*
Mar.2016	**All India Rank-6**, in IEEE Programming League, among over 1200 undergraduate students.	*India*
Mar.2016	**2nd Rank**, CodeWars,a competitive-programming event hosted by IEEE,PEC on CodeChef	*PEC*
Nov.2016	**Research Scholarship**, for Personal Emotional Doctor - Bot	*PEC*
2014	**All India Rank-885**, JEE-Mains, among 1.4 million candidates	*India*
2014	**1st Rank-Opener,PEC**, for **best** JEE-Mains rank among 600 students of the session 2014-2018.	*PEC*
2014	**1 Lakh Scholarship**, from CBSE for 96.4% marks in 12th Boards and 10 CGPA in 10th.	*India*
2014	**Letter of Appreciation**, from HRD Ministry,Govt. of India for 96.4% in CBSE-12th exams	*India*
2011	**Catch Them Young**, Was among the **top-40** students selected from tricity by INFOSYS for 2-week Programming-Basics training on their campus	*INFOSYS*

Skills & Courses

Mathematics	Discrete Structures for Computer Science, Vector Calculus, Fourier Series and Laplace Transform, Operations Research, Mathematics for Machine Learning: Linear Algebra and Multivariate Calculus (Coursera)
Computer Science	Data Structures and Algorithms, Computer Architecture and Organization, OOP, Microprocessor, DBMS, Operating Systems, Computer Networks, Theory of Computation, Artificial Intelligence, Computer Graphics, Mobile Computing, Fastai: Part 1 and Part 2, DeepLearning.ai by Andrew Ng, Deep Learning by Prof. Mitesh Khapra, IIT Madras
Programming & Web	Python, C, C++, Javascript, TypeScript, EcmaScript6, AngularJS, ReactJS, Angular4, Webpack, Django with Python
Machine Learning tools	Pytorch, Pandas, Numpy, ScikitLearn, SciPy, Fastai, Transformers library

Extracurricular Activity

Motivational-Speaker

2014 - PRESENT

- Given more than 10 motivational lectures at schools, college etc to an audience of about 100, talking about *student-life challenges* and *success-hacks*.