

AYAN SENGUPTA

✉ ayan.sengupta007@gmail.com
🌐 <https://github.com/victor7246>

☎ +91 9706318732 📍 Noida, India **in** <https://www.linkedin.com/in/ayans2017/>
🔗 <https://scholar.google.com/citations?user=90EGfboAAAAJhl=en>

PROFESSIONAL EXPERIENCE

UnitedHealth Group (Optum Global Solutions)

Senior Data Scientist

📅 August 2019 – Ongoing 📍 Noida, India

- Working on building NLP solutions to extract structured information from noisy semi-structured scanned clinical charts enabling Optum in saving \$1B+ in risk adjustment business
- Handling several research portfolios on clinical named entity extraction, clinical text summarization, patient health assessment and healthcare graphs

Data Scientist

📅 May 2017 – August 2019

- Built and deployed big-data analytics pipeline to detect patient health risk information for 7M+ member population
- Worked on ML based medical chart scoring and improved false negative rate by 30% leading to reduction of \$70M+ in manual chart coding cost
- Handled various R&D activities on medical fraud detection, voice of customer Analytics and NPS analytics

Corporate accolades received -

- 2021 Optum senior inventor award for contributing towards R&D initiatives
- 2019 innovation and 2018 Q4 mastermind awards for innovation

RESEARCH PROJECTS & PUBLICATIONS

Robust Code-Mixed Representation Learning

📅 Nov 2020 - Dec 2021

- Proposed hierarchical self-attention model for robust representation learning on 6+ Indian code-mixed languages for classification and generative tasks (published in *ACL-IJCNLP 2021*)
- Explored an universal code-mixed encoder for transfer learning and zero-shot learning in code-mixed regime

Hate and Aggression detection from Social Media

📅 Nov 2020 - Apr 2021

- Proposed an unified deep framework for offense detection from Hindi-English code-mixed texts for online social medias. (published in *Neurocomputing journal*)
- Devised a novel causality framework to understand casual dependence among different offensive traits

Framework for Automated Data Denoising

📅 Aug 2020 - Jul 2021

- Proposed a gated-transformer model for detecting and correcting adversarial and machine induced noises from text data (published in *EMNLP 2021*)
- Ideated a multi-modal framework for denoising OCR and speech-to-text noises from real-life data

Joint Topic-Sentiment Model for Short Texts

📅 Mar 2019 – May 2020

- Developed embedding oriented semantic representation learning from texts (published in *AAAI-ICWSM 2021*)
- Developed probabilistic generative model for semi-supervised joint topic-sentiment extraction from texts (published in *SNCS journal*)
- Developed matrix factorization based online topic model for continuous data streams (published in *ICSOC 2020*)

EDUCATION

PhD. in Computer Science

Indraprastha Institute of Information Technology, Delhi

📅 Sep 2021 – ongoing 📍 CGPA - 9.33*/10

Post Graduate Diploma in Business Analytics

IIM Calcutta, IIT Kharagpur, ISI Kolkata

📅 Jun 2015 – May 2017 📍 CGPA - 8.7/10

M.Sc. in Mathematics and Computing

IIT Guwahati

📅 Jul 2013 – Jun 2015 📍 CGPA: 8.8/10

B.Sc. in Mathematics and Computer Science

Chennai Mathematical Institute

📅 Jul 2010 – Jun 2013 📍 CGPA: 6.7/10

OTHER RESEARCH ACTIVITIES

- Program Committee member of ICWSM 2021, ECML-PKDD 2020 and WNUT (EMNLP) 2020
- Reviewer in Journal of Intelligent & Fuzzy Systems
- Co-inventor of 9+ (3 in drafting) US patents

OTHER DATA SCIENCE ACTIVITIES

- Represented ISI Kolkata in *International Data Science Game 2016* finals in Paris and achieved global rank 14 among 100+ teams
- Developer and maintainer of package *jointtsmodel* and *consNLP*
- 2 Bronze medals in Kaggle NLP challenges in 2020. Kaggle Competition Expert
- Won \$80K+ across various data science competitions and gigs in Topcoder between 2017-2021

OTHER ACCOMPLISHMENTS

- Recipient of *National KVPY scholarship* during 2009-2015 for pursuing research (top 200 among 50K+ applicants)
- Cleared *Regional Maths Olympiad* from W.B. region in 2008 (top 50 among 5000+)
- Hold KPMG Six Sigma Green Belt

TECHNICAL SKILLS

- Development, DB

Python R SQL Shell Hive

- ML Frameworks

sklearn PyTorch h2o Sagemaker