Ankit Kumar Pal (Aaditya)

CONTACT INFORMATION

aadityaura.github.io

E-mail: aadityaura@gmail.com
Links: Google Scholar, Github, LinkedIn

RESEARCH INTERESTS

Representation Learning on Graphs, Generative Large Language Models (LLMs), and their applications in Healthcare data, Federated learning, ASR & Audio Analysis

EDUCATION

Babu Banarasi Das University,, Lucknow, India **May 2017**
Bachelor of Technology, Computer Science Engineering

- **Thesis**: Generative Modeling of Music Sequences with LSTM-based RNN Architecture

Anandi Devi S.V.M, Sitapur, India (ADSVM), Sitapur, India **April 2013**
12th - Board of High School and Intermediate Education U.P

- **Major**: Physics, Chemistry and Mathematics

EXPERIENCE

Saama Technologies **May 2018 - Present**
Senior ML Research Engineer

Objective: *Develop Deep Learning/NLP methods and pipelines for clinical data, Lead research projects, and published findings in top ML conferences*

- **Adverse Event Prediction**: *FDA Adverse Event Reporting System (FAERS)* Developed an RNN-LSTM model with Context-Aware Attention to extract pharmacological semantics from clinical notes, achieving 98% F1 score. Optimized character and word embeddings to enrich contextual representation. Enabled automated adverse event detection across 1M records.
- **Trial Plan Optimizer (TPO):** Designed an ML model using one of top-tier biopharmaceutical company's clinical trial data to predict site enrollment. Implemented a Python & Scala AutoML framework with TransmogrifAI. Utilized Categorical Embeddings and tree-based algorithms like XGBoost, LightGBM, and Random Forest to optimize predictions.
- **Unsupervised Medical Monitoring:** Conducted analysis of clinical trial data across SDTM domains to identify patient outliers. Leveraged historical patient data and unsupervised models like Autoencoders, Clustering(e.g. K-Means, DBSCAN), Isolation Forest, and One-Class SVM to optimize outlier detection.
- **DeepMap ML Framework (SDTM Automap):** Developed an ML system to automatically generate CDISC SDTM mappings, incorporating Generative Adversarial Networks, Bidirectional LSTM with PubMed and BERT embeddings, and a 3-layer ELMo architecture for multi-task learning across clinical domains, achieving an average accuracy of 95% in mapping source raw data to SDTM standards.
- **Pharma Graph:** *Predictive Modeling of Drug Interactions using Graph Convolutional Networks* Built a NER model to extract pharmacological relationships from clinical text. Developed a Graph Convolutional Neural Network with attention mechanisms to model drugs as nodes and their interactions as edges, characterizing consequential effects caused by drug pair interactions.
- **Large Language Models for Healthcare Domain** Extracted clinical insights from raw medical documents and PDFs using Retrieval-Augmented Generation, fine-tuned open-source LLMs (e.g., Llama-2, Falcon) using custom instruct-datasets for internal use cases, Developed a Python library for prompt versioning and structured outputs, Generating protocol documents from minimal inputs, and Conducting Research to mitigate LLM hallucinations in the medical domain.

Prescience Decision Solutions, Bengaluru, India **Feb 2018 - May 2018**
Deep Learning Engineer

Objective: *Building a Multidimensional Deep Learning Model to Predict the Bitcoin Price*

- Worked on transfer learning, attention methods, and custom POS-Tag embeddings.
- Created an unofficial Twitter API to get Bitcoin tweets and used it to do LSTM sentiment analysis.
- Added the sentiment analysis as a feature layer in the main model to improve understanding of the data.
- Deployed the code & APIs and built a Chat UI on top of it to interact with the model.

Fliptango Global Solutions, Kerala, India **Dec 2017 – Feb 2018**
Machine Learning Intern

Objective: *Design and implement an ML-driven e-commerce chatbot to optimize user interactions and enhance product recommendations*

- Used TensorFlow to leverage transfer learning and optimize models for specific tasks.
- Added new Commonsense Embeddings from ConceptNet Numberbatch to improve understanding of language.
- Followed BiLSTM-CNN-CRF paper closely to build a named entity recognition model in TensorFlow. Achieved 95% accuracy in the NER model, which was great for pulling out the key entities from user chats.

Selected Publications

Ankit Pal, Muru Selvakumar, Malaikannan Sankarasubbu. Multi-label Text Classification using Attention-based Graph Neural Network. In Proc. ***ICAART, '20***. [Link]

Ankit Pal, Malaikannan Sankarasubbu. Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In ***ACM '21***. [Link]

Ankit Pal. CLIFT: Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain. Poster in ***NeurIPS, '22***. [Link]

Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Proc. ***PMLR '22.*** [Link]

Madhura Josh*, **Ankit Pal***, and Malaikannan Sankarasubbu. Federated learning for healthcare domain - pipeline, applications and challenges. In ***ACM '22.*** [Link].

Ankit Pal. DeepParliament: A Legal domain Benchmark & Dataset for Parliament Bills Prediction. In Proc. ***EMNLP '22.*** [Link]

Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In Proc. ***EMNLP Conll '23.*** [Link]

Service

Reviewed Papers for Springer Nature 2021, IEEE Access 2021, IEEE Access 2022

Technical Skills

- **Programming:** Python, C language, Scala, Rust
- **Mobile and Web Technologies:** HTML, CSS, JavaScript
- **Cloud platforms:** Amazon web services, Google Cloud Platform, and Microsoft Azure
- **Tools:** Jax, Tensorflow, PyTorch, Keras, Scipy, Pandas, Numpy, LaTeX

*equal contribution

Teaching Experience

- **Shala by IIT Bombay:** DL PI-2
  Graph Convolutional Networks for NLP & Knowledge graphs

ML Projects

Covid-19 Question-Answering Bot [2020]

- Extracted keywords and retrieved relevant passages using vector search.
- Ranked top 5 passages for relevance, selecting the top one.
- Summarized chosen passage using the BART model
- Developed APIs and deployed the solution through a Telegram bot.

Image & Product Similarity in E-commerce [2018]

- Transformed product pages into graphs for structural comparison.
- Applied graph isomorphism techniques to identify product similarities.
- Leveraged image vectors to ascertain visual similarity between products.
- Enhanced product recommendation accuracy through combined structural and visual analysis.

Music Generation with LSTM & Double Stacked GRU [2017]

- Transformed MIDI files into encoded matrices for processing.
- Trained both single-layer and double-stacked layer models using LSTM and GRU for music generation.

Voice-Controlled Robotic Arm [2016]

- Constructed a robotic arm with servos, operated by Raspberry Pi on Puppy Linux.
- Integrated a text-to-speech module to translate vocal commands into actionable tasks.
- Enabled the robot to execute diverse actions, like grasping a cup and lifting a ball.
- Secured the second prize in a college technical exhibition for innovation.

Talks

MLOps: The Keystone of Sustainable AI, Coimbatore, India — Jan, 2023
Gradient Optimizers Meetup

Federated Learning & Distributional Shift in Healthcare, Chennai, India — Dec, 2022
Gradient Optimizers Meetup

AI in Law: A New Legal Era, Kangra, India — Oct, 2021
District Court Kangra

Reasoning in LLMs Through Math Word Problems, Chennai, India — Oct, 2020
ML Researchers Meetup

Graphs Neural Networks for NLP, IITB, India — Jul, 2020
Indian Institute of Technology Bombay, Shala

Featured Open-Source Projects

Promptify — **Jan, 2023**
Python and JavaScript — **[Github]**

- A module for prompt engineering and versioning, Enabling users to efficiently utilize the GPT and similar prompt-based models
  to get structured output for various NLP tasks, including NER, QA, Classification, etc
- Github Trending repository

Research Papers Search (Resp) — **Jul 15, 2022**

Python **[Github]**

- A module to Retrieves paper citations from Google Scholar
- Fetches relevant papers by keywords across sources like ACL, ACM, PMLR, etc.

Mix Data types clustering (Mixclu) **Jan 29, 2022**
Python **[Github]**

- Mixclu is an open-source Python package for doing unsupervised mix data types clustering.
- Includes a variety of combination methods including kmeans-onehot, gower distance, umap, etc.

Cough Signal Processing (CSP) **June, 2020**
Python **[Github]**

- Extracts cough features including spectrograms, contiguous segments, and cough events, etc.
- Implements various ML and DL algorithms for respiratory audio analysis tasks including automated cough classification, clustering, anomaly detection, etc.

Unsupervised Toolbox Library (Unbox) **Mar 18, 2020**
Python **[Github]**

- Implements unsupervised NLG methods, focusing on context-aware semantics embeddings.
- Workflow includes question generation/answering, deep clustering, and latent representations etc

Multi-label classification Package (MultiLab) **Nov 26, 2019**
Python **[Github]**

- Implements classical and state-of-the-art models for multi-label classification
- Provides dataset preprocessing, loading, and evaluation metrics