

HARSH AGRAWAL

(857)-313-0855 | Boston, MA | agrawal.har@northeastern.edu | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#) | [Portfolio](#)

EDUCATION

Northeastern University

Boston, MA

Master of Science, Computer Science, **GPA – 3.7**

September 2022 - August 2024

- **Relevant Courses:** Programming Design Paradigm, DBMS, Algorithms, Pattern Recognition and Computer Vision, ML

Narsee Monjee Institute of Management Studies

Mumbai, India

Bachelor Of Technology (Hons.), Computer Engineering, **GPA – 3.75**

July 2018 - August 2022

- **Relevant Courses:** Artificial Intelligence, Image processing, Soft Computing, Natural Language Processing

SKILLS

Languages: Python, Java, C, C++, SQL, R, JavaScript, HTML, CSS

Frameworks: TensorFlow, PyTorch, Scikit Learn, Keras, NumPy, Pandas, OpenCV, Hadoop, Spark, Junit

Tools/IDE: Jupyter, Linux, Git, GCP, AWS, Sage Maker, Data Lake, Glue, Athena, Lambda, Docker, Tableau, MATLAB, MySQL, Snowflake, Firebase, Kubernetes, Apache Kafka, CUDA, Continuous Integration (CI)

Technologies: LLM, Machine Learning, Deep Learning, NLP, Computer Vision, Data Warehousing, Cloud Computing, Gen AI

Publications: 10.1109/CONIT51480.2021.9498561, 10.1109/ICCCNT51525.2021.9579920, 10.1109/ICAIS50930.2021.9395895

PROFESSIONAL EXPERIENCE

BulkMagic

Boston, MA

Machine Learning Engineer

October 2024 – Present

- Led and prototyped a real-time collaborative filtering-based recommender engine for a group-buying platform, enabling collective purchasing power and helping shoppers secure discounts with minimal effort
- Collaborated with cross-functional teams to define data pipelines, performance metrics, and A/B testing frameworks, ensuring streamlined user experiences while optimizing dynamic pricing strategies and platform scalability
- Investigated cutting-edge ML architectures, such as transformer-based encoders and graph-based recommender systems, to personalize deals and reduce time spent on deal-hunting by leveraging collective intelligence

Amazon Robotics

Boston, MA

Data Scientist Co-op

August 2023 – December 2023

- Developed a system to classify and categorize support tickets based on complexity, addressing the issue of ticket backlog by employing custom clustering algorithms on integrated data from multiple sources, using AWS SageMaker and Glue
- Designed a comprehensive downtime monitoring system for robotic arms, using AWS Lambda and Athena to optimize operations, identifying top contributors to downtime, and successfully mapping 60% of downtime occurrences
- Conducted extensive data analysis using AWS Data Lake, SQL, and PostgreSQL to gather and process large datasets and applied ML techniques to solve operational challenges, decreasing downtime for the robotic arm by 15%

PHEME Software Pvt. Ltd

Mumbai, India

AI Developer Intern

April 2021 – August 2021

- Conceptualized, developed, and integrated an AI-based online examination system with anti-cheating features (facial recognition feature using PyTesseract) into LMS platforms, resulting in a 20% reduction in cheating incidents, 40% higher user engagement, and 95% accuracy in detecting misconduct

PROJECTS & RESEARCH EXPERIENCE

Progress Note Understanding: Assessment and Plan Reasoning

May 2024 – August 2024

- Engineered and fine-tuned LLM-based transformer models (BERT, ClinicalBERT, and BiLSTM) to classify relationships in clinical notes, achieving a Macro F1 score of 0.780, with a focus on improving model generalization in healthcare tasks
- Optimized Tiny-ClinicalBERT and Tiny-BioBERT using transformer-layer distillation, aligning the attention maps and hidden states to reduce model size by over 60% while retaining 95% of the original performance
- Designed a high-throughput data preprocessing pipeline using de-identification removal, sentence boundary detection, and dynamic tokenization, coupled with feature extraction (e.g., syntactic dependencies) on the noisy MIMIC-III dataset

Transformative Approaches in EEG Analysis (Detecting Harmful Brain Activity)

January 2024 – May 2024

- Developed a framework using CNNs (EfficientNetB2, MobileNetV3Large, ResNet V2, DenseNet) with TensorFlow and Keras to classify EEG patterns indicative of harmful brain activity, achieving 81.92% accuracy with EfficientNetB2
- Preprocessed EEG and spectrogram data (normalization, log transformation, standardization) using NumPy and Pandas, enhancing model performance and utilizing Kullback-Leibler divergence for probability modeling
- Optimized real-time analysis models by implementing efficient data retrieval and processing pipelines with PySpark and Apache Parquet, improving the speed and accuracy of detecting harmful motor signs in ill patients

Personalized GIF-based Reply Recommendation System

January 2022 – May 2022

- Formulated an approach for predicting relevant GIFs to be used as replies to text messages, resulting in a 45% increase in prediction accuracy using the VINVL transformer as compared to standard OSCAR transformer
- Implemented Twitter API scripts to collect over 1.5M tweets and leveraged LLM-driven text embeddings to match tweets with 115k GIFs (expanded via generative AI), resulting in an accuracy rate of over 80%
- Engineered and built a custom dataset-based collaborative filtering recommendation system that uses sentiment analysis and user characteristics to deliver personalized replies, reducing response time by up to 50%