Harshit Pandey

(617)-785-7063 | Boston, MA | hp2pandey1@gmail.com | linkedin | github/harsh4799 | scholar.google

Education

Northeastern University Boston, MA
Master of Science, Computer Science; ***GPA: 3.94/4.0*** *September 2022 – May 2024*

- **Relevant Coursework** - Natural Language Processing, Machine Learning, Data Mining, Information Retrieval, Algorithms
- **Leadership & Academic Positions:** Student Mentor, Graduate Teaching Assistant (DS:3000), Graduate Research Assistant (SATH Lab)

Savitribai Phule Pune University Pune, India
Bachelor of Engineering, Computer Engineering; ***GPA: 9.08/10.0****; Language Research Group (Core Member)* *July 2017 – July 2021*

Work Experience

MFS Investment Management | **Data Science Intern - Distribution & Sales** Boston, MA
Machine Learning, Natural Language Processing, Large Language Models, Python, SQL *July 2023 - January 2024*

- Collaborated to develop and deploy **ML pipelines**, including data curation, experimentation, modeling, and productionizing.
- Leveraged LDA and **BERT Topic** to categorize and cluster over **1 million** client meeting notes, reducing time spent on analysis.
- Found key buy/redemption influencers, using & **finetuning Financial Bert** for sentiment signal resulting in **11%** increase in r^2 score.
- Engineered a framework for hyperparameter optimization, model selection, & tuning, resulting in **38% faster** experimentation phase.

Northeastern University | **Research Assistant - Machine Learning - SATH Lab** Boston, MA
Python, NLP, Machine Learning, Timeseries Analysis (Changepoint Detection, FFT, etc.), NumPy, Sklearn, Pandas *May 2022 - June 2023*

- Conducted research at the intersection of Health Sciences & Natural Language Processing using Juvenile Abuse Data.
- Studied text messaging data from **29** Juveniles to identify areas of disagreement and investigate the disparities among participants.
- Used **changepoint detection** & **sentiment analysis** with deep learning techniques to identify patterns in the data.
- **Accepted for publication** at the **IEEE's ACII-2023**, & presented at **MIT Media Lab** in September 2023.

Cognizant | **Machine Learning Engineer - Cyber Security Analysis & Prevention** Bangalore, India
Python, Machine Learning, Anomaly Detection, Spark, Oracle SQL, Data Analytics, AWS Sagemaker *February 2021 - July 2022*

- Wrangled **2.3 TB** of **unstructured data logs** using **Spark** for **anomaly detection**, uncovering insights, malicious users & threats.
- Implemented model deployment strategies and established real-time monitoring for optimized threat mitigation with **MLflow**.
- Deployed a responsive **feedback loop**, contributing to **23% reduction** in false positives in models based on **emerging attack patterns.**
- Enhanced cybersecurity measures across **10+ applications**, resulting in an overall **24% reduction** in the risk of cyber attacks.

Projects and Research Publications

Diachronic Word Embeddings: Statistical Insights into Linguistic Evolution | Publication in EMNLP 2021 | Code
Python, NLTK, NLP, Word Embeddings, Statistical Analysis, StreamLit, NumPy, SkLearn, Pandas, Matplotlib

- Utilized various analysis techniques, such as keyword extraction, **trend prediction** based on Productivity metrics, bi-gram tracking, **Semantic Drift analysis**, and similarity monitoring, creating a comprehensive research trend toolkit.
- Analyzed **27,384 abstracts** from the arxiv.cs.CL corpus, providing a thorough exploration of the **Computation & Language domain**.
- Achieved notable success on **GitHub with over 100 stars** & being featured on Papers with Code.

Spotify Podcasts: LLM Powered Document Ranking & Retrieval | Publication in TREC-2020
Large Language Models, Information Retrieval (IR), NLP, Python, NLTK, Pytorch, HuggingFace, NumPy, SkLearn, Pandas, Matplotlib

- Completed an information retrieval project on a large dataset of **100,000+ podcast transcripts** from Spotify.
- Utilized **LLMs** like XLNet with BM25 & RM3 to get top 1k results & developed contextual representation approach for efficient inference.
- **Achieved top 3** ranking among competitors with a high nDCG score of 0.5414.

End-to-End Search Engine from Scratch: Web Scraping, Inverted Indexing, and Advanced Scoring
Information Retrieval (IR), Natural Language Processing, Python, Elasticsearch, Kibana

- Scraped **120K** documents using web crawling techniques & optimized information retrieval across **3 nodes** for efficient merging.
- Implemented inverted indexing & **MapReduce** while using multiprocessing for faster merging. Used Elasticsearch for index storage.
- Experimented with advanced scoring mechanisms such as **BM25**, **language modeling**, & proximity search for improved search results.

Technical Skills

Programming Languages: Python, Java, JavaScript, C++, SQL, NoSQL, Bash, C, CSS, HTML
Tools & Technologies: AWS (EC2, S3, RDS, ECR, SQS, AWS Sagemaker and more), Linux (UNIX), Docker, Git, Jenkins, Excel
Frameworks: PyTorch, MLFlow, Keras, TensorFlow, Flask, NodeJs, Springboot, ReactJs, Angular
Data Science: NLP, A/B testing, Statistics, Classification, Unsupervised Learning, Ensemble, IR, Time Series Analysis, Hypothesis Testing
Big Data & Machine Learning: Spark, Hadoop, MongoDB, MLFlow, AWS Sagemaker, Python (ex. scikit learn, numpy, pandas, matplotlib)

Achievements & Extracurriculars

Best Research Paper Award: Secured **ACM SE 2022's Award** for delivering a standout presentation on a novel and unique solution.
Top 5 Finalists in Oracle Hackathon: Ranked **5 out of 390 teams** in 36-hour hackathon for end-to-end crop recommendation system.
Open Source Contributions: Contributed to HuggingFace's Bigscience Project, DecipticonNLP Library, & Adversarial Deep Learning.
Course Certifications: Earned certifications in AWS Cloud Practitioner, Coursera's Deep Learning Specialization & many other programs.