Shaurya Rohatgi

Washington DC · Twitter · Github · shauryr@gmail.com · +14846325122

Education

Pennsylvania State University - Cohort Fall'17 — State College, PA
PhD - Information Sciences and Technology — August 2017 - May 2023
Thesis: Design and Data Mining Techniques for Large-Scale Scholarly Digital Libraries and Search Engines

Indian Institute of Information Technology and Management — Gwalior, MP, India
Integrated Post Graduate - Information Technology — July 2009 - June 2014
Key Courses: Operating Systems, Data Structures, Design and Analysis of Algorithms

Experience

AllSci — North Bethesda, MD
Applied Scientist — October 2023 - Present

- Goals: Use AI to break down science into its core components - hypothesis, research questions and findings.
- Lead the integration of Large Language Models (LLMs) to facilitate scientific research and enhance analytical reasoning capabilities.
- Tasked with the optimization and implementation of advanced open-source LLMs, including Llama-3-8b and Mistral-7b, tailoring these models to specific institutional needs.
- Successfully engineered a cost-effective strategy that led to a dramatic 90% reduction in OpenAI API expenses, without compromising model performance.
- Build end to end pipelines from raw data, pre-processing, modeling, evaluation to production and deployment.
- Fine-tuned, evaluated and deployed multiple 8 billion parameter LLMs for structured generation at scale

University of Chicago — Chicago, IL
Computational Scientist, Research Computing Centre — July 2023 - September 2023

- Orchestrated the development of smaRT (System Monitoring AI for Request Tracking), integrating NLP and ChatBot technology to streamline service ticketing processes.
- Utilized machine learning techniques to automate and expedite ticket resolution, resulting in a dramatic 80% improvement in operational efficiency.
- Employed data analytics to monitor system performance, delivering insights that informed continuous improvement initiatives.

Allen Institute of Artificial Intelligence — Seattle, WA
Research Intern/Collaborator — May 2021 - December 2022

- Innovated in the field of Science of Science with the creation of a novel academic mentorship prediction dataset, advancing the predictive capabilities of machine learning models.
- Drove the large-scale development of a paper clustering system employing machine learning, significantly contributing to the semantic analysis and categorization of 800+ million academic documents.
- Disseminated research findings through open-source contributions, strengthening the global AI research community.

The Intelligent Information Systems Research Laboratory — State College, PA
Research Assistant — January 2018 - June 2023

- Contributed to the MathSeer project, deploying state-of-the-art machine learning models to augment mathematical formula indexing, retrieval, and ranking, boosting search precision by 40%.
- Led the strategic collection of 30 million open-access documents, implementing a scalable, multithreaded crawler with high-throughput capabilities.
- Championed the deployment and management of CiteSeerX, enhancing the reliability of a major digital library and contributing to its advanced information retrieval systems.

Tata Research — Noida, UP, India
Researcher — December 2014 - July 2017

- Pioneered the development of dialogue-based Natural Language Understanding (NLU) systems, integrating AI chatbots with existing knowledge bases to reduce IT support tickets by 30%.
- Formulated and implemented a novel two-stage clustering algorithm for email categorization, setting a high-performance baseline with a 75% F-Measure.

Publications

For full list of publication please visit : Google Scholar, Semantic Scholar

1. Rohatgi, S., Qin, Y., Aw, B., Unnithan, N., Kan, M. Y. (2023). The ACL OCL Corpus: advancing Open science in Computational Linguistics. arXiv preprint arXiv:2305.14996.
2. Karishma, Z., Rohatgi, S., Puranik, K. S., Wu, J., Giles, C. L. (2023). ACL-Fig: A Dataset for Scientific Figure Classification. arXiv preprint arXiv:2301.12293.
3. Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... Weld, D. S. (2023). The Semantic Scholar Open Data Platform. arXiv e-prints, arXiv-2301.
4. Wu, J., Rohatgi, S., Angadi, M. K., Puranik, K. S., Giles, C. L. (2022, December). Design Considerations for a Sustainable Scholarly Big Data Service. In Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 83-87).
5. **Rohatgi**, S., Downey, D., King, D., Feldman, S. (2022). S2AMP: A High-Coverage Dataset of Scholarly Mentorship Inferred from Publications. arXiv preprint arXiv:2204.10838.
6. Wu, J., **Rohatgi**, S., Keesara, S. R. R., Chhay, J., Kuo, K., Menon, A. M., ... Giles, C. L. (2021, December). Building an Accessible, Usable, Scalable, and Sustainable Service for Scholarly Big Data. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 141-152). IEEE.
7. **Rohatgi**, S., Giles, C. L., Wu, J. (2021, September). What Were People Searching For? A Query Log Analysis of An Academic Search Engine. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 342-343). IEEE.
8. Kandimalla, B., **Rohatgi**, S., Wu, J., Giles, C. L. (2021). Large scale subject category classification of scholarly papers with deep attentive neural networks. Frontiers in research metrics and analytics, 5, 31.
9. **Rohatgi**, S., Wu, J., Giles, C. L. (2021). Ranked List Fusion and Re-ranking with Pre-trained Transformers for ARQMath Lab.
10. **Rohatgi**, S., Karishma, Z., Chhay, J., Keesara, S. R. R., Wu, J., Caragea, C., Giles, C. L. (2020, September). COVIDSeer: Extending the CORD-19 Dataset. In 20th ACM Symposium on Document Engineering, DocEng 2020. Association for Computing Machinery, Inc.
11. **Rohatgi**, S., Wu, J., Giles, C. L. (2020). PSU at CLEF-2020 ARQMath Track: Unsupervised Re-ranking using Pretraining. In CLEF (Working Notes).
12. Zhong, W., **Rohatgi**, S., Wu, J., Giles, C. L., Zanibbi, R. (2020, April). Accelerating Substructure Similarity Search for Formula Retrieval. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 1417, 2020, Proceedings, Part I (pp. 714-727).
13. **Rohatgi**, S., Zhong, W., Zanibbi, R., Wu, J., Giles, C. L. (2019). Query Auto Completion for Math Formula Search. arXiv preprint arXiv:1912.04115.

14. Kim, K., **Rohatgi**, S., Giles, C. L. (2019, November). Hybrid deep pairwise classification for author name disambiguation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2369-2372).

15. **Rohatgi**, S., Zare, M. (2017). DeepNorm-A Deep Learning approach to Text Normalization. arXiv preprint arXiv:1712.06994.

Projects

Experience with Open Source Projects:

- ***Llama-hub***: Engineered an innovative plugin-based architecture for integrating custom data sources with Language Models (LLMs), enhancing the functionality of LLMs through efficient data referencing. Authored an advanced reader module enabling scholarly citation-driven answer generation within LLMs, facilitating academic research with minimal coding effort. The project has garnered significant recognition with over 2.3k stars on GitHub. [code]
- ***Refstudio***: Contributed to the development of a specialized text editor optimized for reference-heavy writing, incorporating Language Model support to streamline the academic writing workflow. Key contributions included the implementation of rewrite functions, conversational interfaces, predictive text completion, intelligent search, and automated citation recommendations, leading to enhanced user productivity. [code]
- ***S2QA: Research Question Answering***: Pioneered a research Q&A tool employing Semantic Scholar and GPT-4 to provide authoritative answers drawn from top-tier research papers. Instrumental in reaching 1,000 user queries on the launch week, the tool leverages state-of-the-art technologies in natural language processing and LLMs to deliver precise information retrieval. [code]

PSU at CLEF'2020 - ARQMath Task: Summer'20
Achieved an outstanding NDCG score, surpassing 94% of participants and setting a new benchmark in the ARQMath Task by leveraging deep learning techniques in both textual and mathematical information retrieval, showcasing significant advancements over existing methods. [slides] [code]

COVIDSeer: Spring'20
Utilized ElasticSearch and Django to architect a high-performance academic search engine focused on COVID-19 literature, implementing advanced indexing techniques on the CORD dataset to facilitate efficient information discovery. [link] [code]

Microsoft Malware Detection Challenge: Fall'18
Conducted a comprehensive exploratory data analysis and utilized Random Forests to predict key factors in malware detection, contributing to proactive cybersecurity measures. [slides] [report]

Deception Detection in Online Dating: Spring'18
Applied machine learning algorithms to analyze the OkCupid dataset, developing a Random Forest classifier that successfully identified deceptive behavior with a high F1 score of 0.94. [slides] [code]

Text Normalization: Fall'17
Crafted a cutting-edge two-stage sequence-to-sequence model utilizing TensorFlow and XGBoost, which achieved top-50 ranking in a Kaggle competition. The model enhances text normalization processes

Skills

Technologies and Frameworks:	SQL, PyTorch, Keras, TensorFlow, Apache Hadoop, ElasticSearch, Scrapy, Heritrix, Docker, Kubernetes, Apache Flink, Apache Airflow, Sagemaker, Streamlit
Operating Systems:	Proficient with Linux and various distributions
Skills:	Expert in Natural Language Processing, Deep Learning, Multimodal Learning, Machine Learning Operations (MLOps), Feature Engineering, Model Development, Validation, and Deployment
Machine Learning Tools:	huggingface, AWS Sagemaker, vLLM, spaCy, NLTK, YARN, Generative AI (including GPT, Llama, and related tools like llama-index, llama-hub, langchain), LLMOps, replicate.ai
Data Analysis:	Proficient in data visualization, pattern recognition, anomaly detection, classification, clustering, and regression techniques
Version Control:	Advanced proficiency with GitHub
Programming Languages:	Python, Java, C, C++, PySpark
Certifications:	
Machine Learning by Stanford University on Coursera:	Certificate
Neural Networks and Deep Learning by deeplearning.ai on Coursera:	Certificate

Awards

Winner AccuWeather Challenge at HackPSU'18 - WeatherOrNot State College, PA
Weather Assisted smart travel suggestions

1st prize winner at HackPSU'17 - FindVisor (IBM Watson Runner Ups) State College, PA
Created a web application which suggests a research adviser for graduate students and ranked professors relevant to them.

Winner Nittany AI Challenge'18 - ProFound: A Professor Search Engine State College, PA
Team Lead - Project was funded for $17,500 and has received support from the Office of Research and College of Medicine, The Pennsylvania State University

Teaching Experience

Instructor - Information Retrieval and Search Engines: IST441
Spring'18, Spring'19, Spring'20, Fall'21
- helped design and was the primary instructor for the lab.
- Students were taught how to crawl the web using ScraPy and build a search engine with the crawled documents using ElasticSearch.
- Worked with graduate students to create custom search engines -PrivaSeer

Academic Service

1. **TheWebConf 2019** (The Web Conference, 2019) : **Subreviewer**
2. **JCDL 2019** (ACM/IEEE Joint Conference on Digital Libraries 2019) : **Subreviewer**
3. **TheWebConf 2020** (The Web Conference, 2020) : **Subreviewer**
4. **CLEF 2020** (Conference and Lab of the Evaluation Forum) : **PC member** (ARQMath - Answer Retrieval for Math Questions)
5. **AI4SG 2021** (International IJCAI Workshop on Artificial Intelligence for Social Good 2021) : **PC member**
6. **CLEF 2021** (Conference and Lab of the Evaluation Forum) : **PC member** (ARQMath-2 - Answer Retrieval for Math Questions)

SELECTED PRESS COVERAGE

1. 2018 Nittany AI Challenge winners provide EdTech solutions, receive $50,000
https://www.psu.edu/news/academics/story/

2018-nittany-ai-challenge-winners-provide-edtech-solutions-receive-50000/
2. HackPSU winning website aims to combat challenge of finding an adviser
https://www.collegian.psu.edu/news/campus/
hackpsu-winning-website-aims-to-combat-challenge-of-finding-an-adviser/article_
4f5dddf6-008a-11e8-b1d6-b3be383b482c.html