

SHAURYA ROHATGI

State College, PA · szr207@psu.edu · [Homepage](#) · +14846325122

EDUCATION

Pennsylvania State University - Cohort Fall'17

State College, PA

Key Courses: Deep Learning, Artificial Intelligence, NLP, Information Retrieval, Data Mining

PhD - Information Sciences and Technology

August, 2017 - Present

Indian Institute of Information Technology and Management

Gwalior, MP, India

Integrated Post Graduate - Information Technology

June, 2014

EXPERIENCE

Allen Institute of Artificial Intelligence

Seattle, WA

Research Intern

May 2021 - Present

- Analysis of Mentorship (May 2021 - April 2022):

Collaborated with the research team at Semantic Scholar (Sergey Feldman and Daniel King) to build a first-of-its-kind dataset and train a model to predict academic mentorship at scale.

1. Crawled sources like ProQuest, Academic Genealogy, and Mathematics Genealogy for the ground truth data (300k samples)

2. Extracted features, trained models, and inferred mentorship for the entire S2 dataset (2 billion links in the authorship network) to answer related science of science questions, such as the characteristics of a mentor that indicate success for a mentee

3. Paper, dataset, and code are available as open-source tools [[paper](#)] [[code](#)]

- Paper Clustering (April 2022 - December 2022):

Working on versioning of papers in Semantic Scholar using machine learning at scale.

1. Clustering and deduplication of 800 million raw academic papers

2. Worked with complex, real-world data

3. Machine learning system was pushed into production

The Intelligent Information Systems Research Laboratory

State College, PA

Research Assistant

January 2018 - Present

1. MathSeer - Indexing and Searching Math in academic documents - 10+ million documents

- Developed novel techniques to index, retrieve, and rank math formulas from academic text - Used pretrained models to improve search ranking by 40%

2. Crawling at scale - Crawled 30M open-access documents. Implemented a custom, multi-threaded, restartable crawler capable of crawling academic PDFs. The crawler completed 30 million downloads in 60 hours.

3. Deploying and maintaining **CiteSeerX** ([Project Lead](#))

- Responsible for keeping the system live: addressed any issues as they arose

- Managed information for 10+ million documents

- Responsible for ML production system related to search and ranking

TCS Research

Noida, UP, India

Researcher

December 2014 - July 2017

1. Dialogue-based systems (NLU) -

- Developed multiple chat agents/bots that retrieved answers from a knowledge base

- Reduced the number of IT support team tickets by 30%.

2. Created a novel 2-stage clustering method for emails, achieving a baseline F-Measure of 75%.

PUBLICATIONS

For full list of publication please visit : [Google Scholar](#), [Semantic Scholar](#)

1. **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.
2. 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.
3. **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.

4. 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.
5. **Rohatgi**, S., Wu, J., Giles, C. L. (2021). Ranked List Fusion and Re-ranking with Pre-trained Transformers for ARQMath Lab.
6. **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.
7. **Rohatgi**, S., Wu, J., Giles, C. L. (2020). PSU at CLEF-2020 ARQMath Track: Unsupervised Re-ranking using Pretraining. In CLEF (Working Notes).
8. 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 14–17, 2020, Proceedings, Part I* (pp. 714-727).
9. **Rohatgi**, S., Zhong, W., Zanibbi, R., Wu, J., Giles, C. L. (2019). Query Auto Completion for Math Formula Search. arXiv preprint arXiv:1912.04115.
10. 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).
11. **Rohatgi**, S., Zare, M. (2017). DeepNorm-A Deep Learning approach to Text Normalization. arXiv preprint arXiv:1712.06994.

RESEARCH PROJECTS

Research Question Answering: Spring'23

- Allows users to enter a question and uses the power of Semantic Scholar and GPT-3 to generate an answer based on the content of top research papers.
- Received 1000 queries on the first day of release [[code](#)][[demo](#)]

PSU at CLEF'2020 - ARQMath Task: Summer'20

Achieved an NDCG' score of 0.31, outperforming 94% of the participants and even surpassing the state-of-the-art in some categories. [[slides](#)][[code](#)]

COVIDSeer: Spring'20

Models and technologies used: *ElasticSearch and Django*

Built a search engine for academic articles related to COVID-19 using the CORD dataset. [[link](#)][[code](#)]

Microsoft Malware Detection Challenge: Fall'18

Conducted exploratory data analysis and used random forests to predict and understand the most important factors in identifying malware-affected computers. [[slides](#)][[report](#)]

Deception Detection in Online Dating: Spring'18

Identified characteristics of people on dating apps who are in relationships and cheating on their partners. Analyzed the OkCupid dataset and used a random forest to identify. Our method reported an F1 score of 0.94. [[slides](#)][[code](#)]

Text Normalization: Fall'17

Models and technologies used: *Tensorflow, XGBoost, Seq2Seq model*

Developed a novel two-stage seq2seq model. The prediction of the classification algorithm is used by the sequence-to-sequence model to predict the normalized text. Ranked among the top 50 in the Kaggle Competition. [[code](#)][[slides](#)]

Fake News Detection: Fall'17

Models and technologies used: *Keras, Tensorflow, Spacy, word2vec*

Detected unreliable news and trained and reported results on the dataset provided by the AICS'18 Workshop challenge. Our deep learning model reported an accuracy of 93%, significantly better than the baseline by 7 absolute percentage points. [[report](#)][[code](#)][[slides](#)]

Math Information Retrieval using CNNs:

As part of my thesis research, we encode math formula images using convolutional autoencoders. These encodings are then used for math information retrieval. [\[code\]](#)

AWARDS

Winner, AccuWeather Challenge at HackPSU'18 - WeatherOrNot State College, PA
Developed weather-assisted smart travel suggestions.

1st Prize Winner, HackPSU'17 - FindVisor (IBM Watson Runner-Up) State College, PA
Created a web application that suggests research advisers for graduate students and ranks 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 received support from the Office of Research and College of Medicine, The Pennsylvania State University.

SKILLS

Packages, OS, Skills and frameworks: SQL, PyTorch, Keras, Tensorflow, Apache Hadoop, ElasticSearch, Linux, Scrapy, Heritrix, spaCy, nltk, distributed system architectures, MATLAB, Sagemaker, Docker, Flink Kubernetes, Natural Language Processing, Deep Learning, Multimodal Deep Learning, Large Language Models, YARN, classification, clustering, regression, visualize patterns, anomalies, airflow feature engineering, model development, model validation and deployment, GitHub, Generative AI, GPT-3

Programming Languages: Python, Java, C, C++, pyspark

Machine Learning by Stanford University on Coursera: [Certificate](#)

Neural Networks and Deep Learning by deeplearning.ai on Coursera: [Certificate](#)

TEACHING EXPERIENCE

Instructor - Information Retrieval and Search Engines: IST441

Spring'18, Spring'19, Spring'20, Fall'21

- Assisted with designing the course and served as the primary instructor for the lab.
- Taught students how to crawl the web using Scrapy and build a search engine utilizing the crawled documents through Elasticsearch.
- Collaborated with graduate students to develop custom search engines, such as [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_4f5ddd6-008a-11e8-b1d6-b3be383b482c.html