

SHAURYA ROHATGI

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

RESEARCH INTERESTS

Natural Language Processing, Information Retrieval and Machine Learning

EDUCATION

Pennsylvania State University - Cohort Fall'17

State College, PA

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

GPA: 3.56/4.0

PhD - Information Sciences and Technology

August, 2017 - Present

Indian Institute of Information Technology and Management

Gwalior, MP, India

Key Courses: Operating Systems, Data Structures,

Design and Analysis of Algorithms

GPA: 6.96/10

Integrated Post Graduate - Information Technology

June, 2014

EXPERIENCE

The Intelligent Information Systems Research Laboratory

State College, PA

Research Assistant

January, 2018 — Present

1. MathSeer - Indexing and Searching Math in academic documents at Scale (Current Research Focus)
2. Crawling at scale - Crawled 30M open-access documents
3. Deploying and maintaining **CiteSeerX** (Project Lead)

TCS Research

Noida, UP, India

Researcher

December 2014 — July 2017

1. Dialogue based systems (NLU)
 - worked on multiple chat agents/bots' answers were retrieved from RDF
 - reduced the number of tickets of the IT support team by 30%.
2. Created a novel 2 stage clustering method for the emails of which reported an F-Measure of 75%.

PUBLICATIONS

- **Rohatgi, S.**, Wu, J., Giles, C. L. (2020). PSU at CLEF-2020 ARQMath Track: Unsupervised Re-ranking using Pretraining. (to appear in CLEF workshop)

- **Rohatgi, S.**, Karishma, Z., Chhay, J., Keesara, S. R. R., Wu, J., Caragea, C., Giles, C. L. (2020). COVIDSeer: Filling missing pieces in the COVID-19 dataset.

- Mansouri, B., **Rohatgi, S.**, Oard, D. W., Wu, J., Giles, C. L., Zanibbi, R. (2019, September). Tangent-CFT: An Embedding Model for Mathematical Formulas. In Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (pp. 11-18). ACM.

- Wu, J., Kandimalla, B., **Rohatgi, S.**, Sefid, A., Mao, J., Giles, C. L. (2018, December). CiteSeerX-2018: A Cleansed Multidisciplinary Scholarly Big Dataset. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5465-5467). IEEE.

- **Rohatgi, Shaurya** and Zare, Maryam . "DeepNorm-A Deep Learning Approach to Text Normalization." arXiv preprint arXiv:1712.06994 (2017)..

- Patidar, M., **Rohatgi, S.**, Chaudhary, A., Singh, M. P., Agarwal, P., Shroff, G. (2016, December). Activity Detection from Email Meta-Data Clustering. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 568-575). IEEE.

RESEARCH PROJECTS

PSU at CLEF'2020 - ARQMath Task:

Summer'20

Achieved a NDCG' score of 0.31 which was higher than 94% of the participants and even beats the state-of-the-art in some categories. [\[slides\]](#)[\[code\]](#)

COVIDSeer:

Spring'20

Models and Technologies used - *ElasticSearch and Django*

Build a Search Engine for academic articles related to COVID-19 using the

CORD dataset.[\[link\]](#)[\[code\]](#)

Microsoft Malware Detection Challenge: Fall'18

Did exploratory data analysis of data and used random forests to predict and understand what factors are most important in identifying malware affected computers. [\[slides\]](#)[\[report\]](#)

Deception Detection in Online Dating: Spring'18

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

Text Normalization: Fall'17

Models and Technologies used - *Tensorflow, XGBoost, Seq2Seq model*

Prediction of the classification algorithm is used by our sequence-to-sequence model to predict the normalized text of the input token. 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 we train and report our results on the dataset provided by the AICS'18 Workshop challenge. Our Deep Learning model reports an accuracy of 93% which is better than the baseline (86%). [\[report\]](#)[\[code\]](#)[\[slides\]](#)

Math Information Retrieval using CNNs: As a 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
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 ([homepage](#))

SKILLS

Packages, OS and frameworks: PyTorch, Keras, Tensorflow, Apache Hadoop, ElasticSearch, Linux, Scrapy, Heritrix

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

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

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

TEACHING EXPERIENCE

Instructor - IST 441 Spring'18

- helped design and was the primary instructor for the lab

[Information Retrieval and Search Engines.](#)

- Students were taught how to crawl the web using ScraPy and build a search engine with the crawled documents using ElasticSearch.