

Saurav Jha

Contact Information	Machine Learning Engineer Cognitive Computing Team Factset Research Systems, Hyderabad, India 500032	phone: +91-89-5350 0973 e-mail: mail@sauravjha.com.np web: www.sauravjha.com.np
Research Interests	I am broadly interested in Machine Learning, Natural Language Processing and Computer Vision. My recent works include multi-task morphology learning, neural machine translation for low-resource languages (LRLs), cartoon face recognition, scientific document summarization, and topic modeling.	
Education	Motilal Nehru National Institute of Technology , Prayagraj, India B. Tech., Computer Science and Engineering (7.53 / 10.0) Coursework: Genetic Algorithm (CS1736), Data Ware Housing and Mining (CS1743), Scientific Computing (CS1602), Database Management Systems (CS1605), Operation Research (CS1505), Cryptography (CS1506), Operating Systems (CS1502), Computer Networks (CS1503), Automata Theory (CS1404), Analysis of Algorithms (CS 1401), Graph Theory and Combinatorics (CS 1402). Thesis supervisor: Dr Suneeta Agarwal	2018
	United Academy , Lalitpur, Nepal 10+2, High School Science (with Distinction)	2013
	Happyland Secondary School , Rajbiraj, Nepal Secondary School (with Distinction)	2011
Employment	Factset Research Systems Inc. , India Machine Learning Engineer Worked on Callstreet subject tagging, keyword extraction, and topic modeling projects.	Jun 2018 - Present
Research Experience	Indian Institute of Technology, Banaras Hindu University , Varanasi, India Research Intern, NLPR Lab Worked on Multi-task morphology learning for Hindi and Urdu, Neural machine translation for low-resource Indian languages, and Scientific Document Summarization in Anil Kumar Singh's group	Jun 2017 - May 2018
	Motilal Nehru National Institute of Technology , Prayagraj, India Undergraduate Research Intern, Image Processing Lab Worked on Cartoon face detection and recognition, and English OCR project in Suneeta Agarwal's group.	Oct 2017 - Apr 2018
Projects	Multi-Task Neural Morphological Analysis <ul style="list-style-type: none">Proposed the Multi-task Deep Morphological Analyzer (MT-DMA) to jointly predict POS, Gender, TAM, Case, Number, Person, TAM, and Lemma of Hindi and Urdu words.Incorporated tag-specific phonological features optimized through Multi-objective GA to establish a state-of-art accuracy score. Web API available here.	Mar 2018 - Oct 2018
	Word Transduction for OOV words in LRLs <ul style="list-style-type: none">Adapted benchmark attention-based variants to transliterate the cognate pairs among OOV words for Hindi-to-Bhojpuri and Hindi-to-Bangla machine translation.Devised an algorithm to extract character level embeddings from pre-trained FastText embeddings besides suggesting methods to create artificial parallel corpora for such languages.	Oct 2017 - Jun 2018
	Scientific Document Summarization Shared Task <ul style="list-style-type: none">Carried out a binary classification of <i>[Citation, Reference]</i> pairs using 1-D CNN to identify candidate sentences contributing to the summary of the referenced paper.Achieved a state-of-art F1 score of 0.5558 on problem 1A of the CL-SciSumm 2017 shared task.	Jun 2017 - Aug 2017

Cartoon Face Detection and Recognition

Oct 2017 - Apr 2018

- Implemented an MTCNN model based on transfer learning to detect cartoon faces with a True Positive Rate of ~78%.
- Proposed a facial key-point feature based CNN architecture for Cartoon face recognition, depicting its on-par potentiality with Inception v3 + SVM and releasing an annotated data set.

Gender Classification of Blog Authors

Jul 2016 - Mar 2017

- Worked on incorporating variable length character sequence patterns to the then state-of-art feature set.

Case sensitive OCR for English Alphabet

Jan 2017 - Apr 2017

- Extracted an 18-D glyph feature vector for predicting case-sensitive English letters with ~92% acc. using ensemble of Logistic Regression, Gradient Boosting and SVM classifiers.

Scholarly works [Google Scholar Profile [here.](#)]

B. Tech. Thesis

Saurav Jha, Nikhil Agarwal, Suneeta Agarwal. *Bringing Cartoons to Life: Towards Improved Cartoon Face Detection and Recognition Systems.* [[Paper](#)][[Code](#)]

Journal Pre-prints

1. Saurav Jha, Akhilesh Sudhakar, Anil Kumar Singh. *Multi-Task Deep Morphological Analyzer: Context Aw-are Joint Morphological Tagging and Lemma Prediction.* [[Paper](#)][[Code](#)]
2. Saurav Jha, Akhilesh Sudhakar, Anil Kumar Singh. *Neural Machine Translation based Word Transduction Mechanisms for Low-Resource Languages.* [[Paper](#)][[Code](#)]

Refereed Conference Papers

1. Saurav Jha, Aanchal Chaurasia, Akhilesh Sudhakar and Anil Kumar Singh. **Reference Scope Identification for Citances using Convolutional Neural Networks.** *14th International Conference on Natural Language Processing (ICON 2017).* [[Paper](#)][[Code](#)]
2. Vijay Prakash Dwivedi, Deepak Kumar Singh, Saurav Jha and Ranvijay. **Gender Classification of Blog Authors: With Feature Engineering and Deep Learning Using LSTM Networks.** *9th International Conference on Advanced Computing (ICoAC).* [[Paper](#)][[Code](#)]

Mini-projects

Educational blogs

- Visualizing the H-index patterns of Scimago Journal Ranking indicators with Pandas. [[Link](#)]
- Retinal fundus image segmentation into background, atrophy and optic disk classes with OpenCV. [[Link](#)]

Data Analysis

- Experimented with one-shot learning using Siamese Manhattan LSTMs to identify citing sentences on the imbalanced dataset from CL-SciSumm 2017. [[Code](#)]
- Built a binary classification model using voting-based ensemble for prediction subscription of a client to a product based on marketing campaign information. [[Code](#)]

Networking

- Set up and monitored a forward proxy server using Squid v3.5 for the academic year 2017/18 at the Computer Center, MNNIT Allahabad.

Technical Skills

1. Languages: **Advanced** – Python, C++, C | **Intermediate** – Java, MySQL, CSS3.
2. Frameworks: Keras, Tensorflow, PyTorch, scikit-learn, OpenCV, Matplotlib, Pandas, Numpy, Flask.
3. Platforms: Amazon Sagemaker, LaTeX, PyCharm.

Test Score **TOEFL-iBT : 107/120**

Audited Courses 1. CS 294: Deep Reinforcement Learning, Fall 2017, UC Berkeley.
2. Deep Learning for NLP at Oxford, DeepMind 2017.
3. Machine Learning from Stanford University, Coursera.

Position of Duties 1. Guided two summer interns (Harshika Arya and Soumil Mandal) on CoNLL-SIGMORPHON task.
2. Supervisor at Hack-36, 2018, the first 36-hour annual hackathon of MNNIT Allahabad.
3. Member of the Content Team at Renaissance 2017, the annual e-summit of MNNIT Allahabad.
4. Member of the Marketing and Publicity panel at Avishkar 2014, the annual tech fest of MNNIT Allahabad.