Saurav **Saha**

Deep Neural Networks · Machine Learning · Python · Data Science · Artificial Intelligence

R-18, Boys Hostel 3, NIT Mizoram, Aizawl PIN : 796025, India

(+91) 9402508957 | contact.srvsaha@gmail.com | SRvSaha | srvsaha

"Be the change that you want to see in the world"

Education

National Institute of Technology Mizoram (NIT Mizoram)

Aizawl, Mizoram, India

B.Tech. in Computer Science and Engineering — ***Aug. 2014 - June. 2018***

- Secured a CGPA of **9.60/10**
- Departmental **Gold Medalist** for scoring highest CGPA amongst all students of Department of CSE.

S.R. Public Sr. Sec. School, Kota, Rajasthan (CBSE)

Kota, Rajasthan, India

High School — ***Apr. 2012 - May. 2013***

- Secured **93%** overall and ranked amongst top 5% in India
- 94% Maths, 92% Chemistry

Green Valley English School, Mankachar, Assam (SEBA)

Mankachar, Assam, India

Secondary School — ***Jan. 1999 - May. 2012***

- Secured **91.67%** in HSLC Examination, ranked **2nd in District** and **18th Rank Holder in State**
- 99% Science, 98% Advanced Maths, 94%- Maths

Experience

B.Tech Final Year Project, Department of CSE, NIT Mizoram

NIT Mizoram, India

Guide : Mr. Sandeep Dash, Department of CSE, NIT Mizoram — ***Aug. 2017 - May 2018***

- **Development of Automatic Image Captioning System.**
- **Field** : Deep Learning, Natural Language Processing, Computer Vision
- Researched on the idea of generating embeddings for multiple modalities and projecting them into same representation space.
- Used transfer learning to initialize activations of ImageNet for images and GloVe embeddings for captions.
- Used ResNet50 CNN architecture for generating features of images and GRU for generating caption embeddings.
- The system achieved a **BLEU score** of **38.2** on test dataset and acquires **accuracy of 95%** for assigning correct caption for an image.
- The developed system is useful in either way retrieval as in related images finding from a given image, caption generation from an image, related images retrieval from caption.
- **Publication** : Paper published in Springer LNCS as a part of **CICLing 2018**
- **Project Link:** *Automated Image Captioning System*

DAAD-WISE Scholar, Summer Research Intern, Universität Bremen, Germany

Bremen, Germany

Guide : Prof. Rainer Malaka, HEAD of DIGITAL MEDIA GROUP, TZI — ***Jun. 2017 - Aug. 2017***

- **Development of an Artificial Episodic Memory model for Robots for Mastering Everyday Activities.**
- **Field** : Natural Language Processing, Knowledge Representation, Artificial Intelligence
- Researched on developing Artificial Episodic Memory using Ontology as Knowledge Base in Project EASE.
- Developed a Recursive Crawler from scratch to scrape kitchenstories.io website unofficially to prepare the dataset.
- Developed a pipeline that turns natural language instructions into individual concept sets corresponding to a given ontology.
- Developed an efficient discourse memory model from scratch which traces the occurrences of individual concepts over the entire set of instructions that encompass the specific everyday activity at hand.
- **Project Link:** *Project EASE*

Data Science Intern, IIT Mandi, Himachal Pradesh

IIT Mandi

Guide : Prof. (Dr.) Varun Dutt, IIT Mandi — ***Dec. 2016 - Jan. 2017***

- **Machine Learning and Data Mining project using Big Data in Health-Care**
- **Field** : Data Mining, Big Data, Machine Learning
- Developed a predictive model to identify from EMR Datasets which patient is likely to buy which medicine using Machine Learning.
- Developed a predictive model for determination of Frequent/Infrequent buyer given the attributes of the patient.
- Reduced the time required for training the system by using Weka-Parallel, a parallel computing architecture.
- Built majority voted ensemble for binary-class and ternary-class classification task from scratch.
- Tuned the performance of Decision Tree ML Algorithm by hyperparameter optimization using GridSearch Algorithm.
- **Project Link:** *Big Data Analytics in Health-Care*

Paraphrase Detection in India Languages, FIRE-ISI 2016 *NIT Mizoram, Aizawl*

Guide : Dr. Partha Pakray, HoD CSE, NIT Mizoram, Mr. Sandeeep Dash, NIT Mizoram *Jul. 2016 - Sept. 2016*

- **Field** : Machine Learning, Textual Semantic Similarity
- Our system NLP-NITMZ is based on three features: Unigram Matching Ratio, Levenshtein Ratio and Cosine Similarity using Vector Space Model.
- Built two classifiers which can tag paraphrases, non-paraphrases and semi-phrases in Indian Languages, namely Hindi, Malayalam, Punjabi and Tamil. Our classifiers are voted ensambles built on the top of Naive Bayes, Support Vector Machines, Random Forest, Logistic Regression, J48 Machine learning algorithms and gives **95%+** accuracy in the Train Set. In Test Set, we got **91.55%** in Hindi, **83.44%** in Malayalam, **94.20%** in Punjabi and **83.44%** accuracy in Tamil.
- For Machine Learning portion we have used **Ensembled Model** to predict the class.
- **Publication** : Paper published on **8th meeting of Forum for Information Retrieval Evaluation (FIRE 2016)**
- **Link:** *CEUR-WS.org/Vol-1737, Pages 256-259*

Winter Research Intern, Jadavpur University, Kolkata *Kolkata, India*

Guide : Prof. (Dr.) Dipankar Das, Jadavpur University *Dec. 2015 - Jan. 2016*

- **Phrase Extraction from English Sentences for Clausal Identification**
- **Field** : Information Extraction, Text Mining, Data Structures, Algorithms
- The system built on the top of Stanford Parser and NLTK can detect various type of Phrases and can separate them automatically which can be used to extract Clauses from texts.
- Developed a recursive algorithm based on stack data structure which keep track of the start and end of phrases within Phrases. The task is recursively solved to extract the phrases along with their type.

Technical skills

Languages : Python, C/C++, JAVA, JSON, SQL. LaTeX
Tools/Frameworks : Stanford Core NLP, Protégé, WEKA, NLTK, Keras, MySQL, NumPy, Scikit-Learn, kdb+ database, git, gdb, Apache Hadoop, Python unittest, Markdown, BeautifulSoup, PyQt
Platform : Linux, Windows, Macintosh

Additional Relevant Courses

Online/MOOC

2017	**Mining of Massive Datasets**, Prof. Jeff Ullman, Prof. Jure Leskovec, Prof. Anand Rajaraman	*Stanford University*
2017	**Introduction to Hadoop and MapReduce**, Sarah Sproehnle, Ian Wrigley, Gundega Dekena	*Cloudera*
2016	**Introduction to Machine Learning**, Prof. Andrew Ng	*Stanford University*

GIAN Courses

2017	**Deep Learning for Natural Language Processing**, Dr. Benoit Favre, Aix-Marseille University(AMU)	*France*
2016	**Natural Language Processing & Sentiment Analysis**, Prof. Alexander Gelbukh, Instituto Politécnico Nacional (IPN)	*Mexico*

Achievements & Awards

Research

2017	**Awarded DAAD-WISE Scholarship**, Research Internship in Universität Bremen, Germany	*DAAD, Germany*
2016	**3rd Place**, Detecting Paraphrases in Indian Languages (DPIL), FIRE'16. Overall : 36 teams	*FIRE, ISI Kolkata*
2016	**Scored an absolute 10/10 grade** , Natural Language Processing & Sentiment Analysis	*GIAN, NIT Mizoram*

References

Prof. (Dr.) Rainer Malaka
Research Professor & Head
Digital Media Lab
Universität Bremen, Germany
Bremen - 28359, Germany
malaka@tzi.de
http://dm.tzi.de/staff/malaka/

Dr. Varun Dutt
Applied Cognitive Science Lab
School of Computing and Electrical Engineering
Indian Institute of Technology (IIT) Mandi
Kamand, Himachal Pradesh, India – 175 005
varun@iitmandi.ac.in
http://faculty.iitmandi.ac.in/~varun/