# Saurav **Saha**

Cognitive Science · Machine Learning · Python · Data Science · Artificial Intelligence

*R-17, Boys Hostel 3, NIT Mizoram, Aizawl PIN : 796001, India*

(+91) 9402508957 | contact.srvsaha@gmail.com | www.sauravsaha.in | 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, Pre-Final Year — ***Aug. 2014 - PRESENT***

- CGPA of **9.57/10** (till Nov 2016, 5th Semester)
- Consistenly maintaining **department rank 1** over past two years

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

High School — ***Apr. 2012 - May. 2014***

- Secured **93%** overall and ranked amongst top 5% in India
- 95% Maths, 95% 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

## Conference/Workshop

2014 **SciPy India**, **International Conference** on Python for Scientific Computing organised by FOSSEE — *IIT Bombay*

2014 **NNSC**, **National Workshop** on Network Security & Implementation organised by Network Bulls — *IIT Bombay*

## Experience

**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 which could identify from EMR Datasets which patient is likely to buy which medicine using Machine Learning Techniques.
- Developed a predictive model for determination of Frequent/Infrequent buyer given the attributes of the patient.
- Case study and analyses of various databases like MongoDB, Cassandra, HBase, kdb+ to understand why kdb+ is better in handling realtime big data.
- Reduced the time required for training the system by using Weka-Parallel, a parallel computing architecture.
- Analyses of handling Big Data in Hadoop vs Spark for Machine Learning
- Built majority voted ensemble for binary-class and ternary-class classification task from scratch.
- Tuned the performance of Decision Tree ML Algorithm by hyperparameter tuning using GridSearch Algorithm.
- **Tools** : Python, Excel, Apache Hadoop, Weka, Weka-Parallel, kdb+

**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 **Probabilistic neural network(PNN)** to predict the class.
- **Tools** : Python, JAVA, WEKA, MATLAB, XML
- **Publication** : Paper published on **8th meeting of Forum for Information Retrieval Evaluation (FIRE 2016), CEUR-WS.org/Vol-1737, Pages 256-259**

**QA4FAQ - Question Answering for Frequently Asked Questions, EVALITA-2016** *NIT Mizoram, Aizawl*

Guide : Dr. Partha Pakray, NIT Mizoram, Head of NLP-NITMZ Research Team ***Jul. 2016 - Sept. 2016***

- **Field** : Information Extraction, Information Retrieval, Text Mining
- Since searching within the Frequently Asked Questions (FAQ) page of a web site is a critical task: customers might feel overloaded by many irrelevant questions and become frustrated due to the difficulty in finding the FAQ suitable for their problems.
- Developed a search-engine which can effectively retrieve a list of most relevant FAQs and corresponding answers related to the query issued by the user. Used Combinatorics approach for query and by rating each result fetched on a scale like 3,2,1, the most relevant one is shown.
- Our system can effectively give **97%** relevant search results based on the queries of the user which is much better than any prevalent IR methodologies.
- **Tools** : Python, JAVA, Nutch, Apache Tomcat, Italian Stop-word Corpus Building, Combinatorics, Page Rating & Ranking Algorithms
- **Publication** : Paper published on **3rd Italian Conference on Computational Linguistics (CLiC-it 2016)**

**Information Extraction from Microblogs(Twitter) Posted during Disasters, FIRE-ISI 2016** *NIT Mizoram, Aizawl*

Guide : Dr. Partha Pakray, NIT Mizoram, Head of NLP-NITMZ Research Team ***Jul. 2016 - Sept. 2016***

- **Field** : Social Media Code Mixing, Information Extraction, Information Retrieval
- Built a system that can deal with the noisy nature of microblogs(**TWITTER**) which are very small (at most 140 characters) and often written informally, using abbreviations, colloquial terms, etc, and effective developed IR methodologies for extracting important information from microblogs posted during disasters.
- Our system can effectively show the most relevant results related to each topic(Tweets) with high precision Twitter data.
- **Tools** : Python, JAVA, Nutch, Apache Tomcat, Word2Vec, Page Ranking Algorithms, Twitter API, JSON, TF-IDF, Skip-gram, Continuous Bag of Words(CBOW)

**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, Data 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.
- **Tools** : Stanford Parser, NLTK, Python, JAVA

## Technical skills

**Languages :** Python(Primary), C, C++, JAVA, JSON, XML, Shell Script, SQL, PHP
**Tools/Frameworks :** WEKA, NLTK, Stanford Parser, Nutch, Apache Tomcat, kdb+ database, Oracle Database, git, vim, gdb, Sublime Text 3, Scikit-Learn, NumPy, Keras, Sphinx, reStructuredText, Apache Hadoop, Microsoft Excel, Python unittest, LaTeX, Markdown
**API :** Twitter API
**Platform :** Linux(Primary), Windows

## 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* |
| 2016 | **Artificial Intelligence**, Prof. Patrick Winston | *MIT* |
| 2015 | **Data Structures and Algorithms**, Dr. Naveen Garg | *IIT Delhi* |

### 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* |
| 2016 | **Introduction to Robot Operating System**, Prof. David Pinto Avendaño, Benemérita Universidad Autónoma de Puebla (BUAP) | *Puebla, Mexico* |

## Achievements & Awards

### Research

| | | |
|---|---|---|
| 2016 | **3rd Place**, Detecting Paraphrases in Indian Languages (DPIL), FIRE'16. Overall : 36 teams | *FIRE, ISI Kolkata* |
| 2016 | **5th Place**, Information Extraction from Microblogs Posted during Disasters, FIRE'16 | *FIRE, ISI Kolkata* |

### Scholastic

| | | |
|---|---|---|
| 2016 | **Scored an absolute 10/10 grade** , Natural Language Processing & Sentiment Analysis | *GIAN, NIT Mizoram* |
| 2016 | **Scored an absolute 10/10 grade** , Introduction to Robot Operating System | *GIAN, NIT Mizoram* |
| 2016 | **Scored an absolute 10/10 SGPA**, Highest till date in CSE Dept, Spring Semester 2016 | *NIT Mizoram* |
| 2015 | **Science Olympiad**, ALL MIZORAM RANK 4th in "MANTHAN", regional SCIENCE OLYMPIAD | *NIT Mizoram* |
| 2014 | **All India Topper**, Secured 99/100 in Physical Education in CBSE 2014, subject merit highest | *CBSE, India* |
| 2014 | **JEE MAIN**, Amongst the top **2.5%** in India, out of 1.4 million appeared candidates | *India* |
| 2013 | **All India Rank 6th**, NATIONAL SCIENCE CONCOURS organised by Maxscore in 2013 | *Gurgaon, India* |

### Extra-Curricular

| | | |
|---|---|---|
| 2016 | **Gold Medallist**, Won the Gold Medal, best Badminton Player of NIT Mizoram in Shaurya'16 | *NIT Mizoram* |
| 2015 | **3rd Prize**, FASTEST RUBIK's CUBE solving competition in Anunaad'15 | *NIT Mizoram* |
| 2015 | **2nd Prize**, Counter Strike, Arcadia 2015 at ANUNAAD, the annual Techno-Cultural Festival | *NIT Mizoram* |

## Positions of Responsibility

| | | |
|---|---|---|
| 2015 - Present | **Teaching Assistant**, Mentoring a group of 20 students giving hands-on exposure to C/C++ Programming and Competitive Programming | *NIT Mizoram* |
| 2015 | **Organiser**, Code Warrior, annual CODING contest of NIT Mizoram at ANUNAAD 2015. | *NIT Mizoram* |
| 2014 | **Senior Managerial Team Member**, Head of 31 NITs in "YUGMA", annual pan NIT magazine | *Inter NIT* |
| 2014 | **Founder**, Founder of HackerRank Club, NIT Mizoram (Coding Club of NIT Mizoram). | *NIT Mizoram* |
| 2014 | **Organiser**, Organised "WEB-o-TIC", the online marketing & publicity events in ANUNAAD'15. | *NIT Mizoram* |
| 2014 | **Campus Ambassador**, Campus Ambassador for Techniche, IIT Guwahati, Seismech, IIT Guwahati, Zigsaw Consultancy Services, HackerRank, PTBN - the Inter NIT Network | *Pan India* |
| 2014 | **Technical Co-ordinator**, Co-ordinated the Techno-Cultural Anunaad'15, responsible for Technical Events Management | *NIT Mizoram* |

## References


**Dr. Partha Pakray**
Head of the Department
Assistant Professor
International Co-operation Officer
Principal Investigator
Department of Computer Science & Engineering
National Institute of Technology (NIT) Mizoram
Aizawl - 796012 , India
+91 8259065018
parthapakray@gmail.com
www.parthapakray.com

**Dr. Dipankar Das**
Assistant Professor
Department of Computer Science & Engineering
Jadavpur University
Kolkata-700032
India
dipankar.dipnil2005@gmail.com
www.dasdipankar.com

**Prof. (Dr.) Alexander Gelbukh**
Research Professor & Head
Natural Language Lab
Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN) Mexico
Mexico City - 07738, Mexico
(+52 1) 55 1810-4587
gelbukh@gelbukh.com
www.gelbukh.com

**Dr. Varun Dutt**
Applied Cognitive Science Lab
School of Computing and Electrical Engineering
Assistant Professor
Indian Institute of Technology (IIT) Mandi
Kamand, Himachal Pradesh, India – 175 005
varun@iitmandi.ac.in
+91-1905-267041
http://faculty.iitmandi.ac.in/ varun/