Mihir Shekhar

Email : mihirshekharcse2010@gmail.com
https://researchweb.iiit.ac.in/ mihir.shekhar/ Mobile : +91 9581826727

Summary

A computer science professional pursuing Ph.D in Data Mining at International Institute of Information Technology, Hyderabad. 5+ years of work experience in machine learning and deep learning.

Skills

- Experience in machine learning and deep learning with applications in NLP (Machine Translation, Chatbot, Entity Identification and Linking) and Health Informatics (Patient Cohort Generation).
- Experience in unsupervised, semi-supervised and weak learning.
- Scikit-learn, NLTK, Stanford NLP, Weka, Elki, Lucene, Keras, PyTorch, D3.js, matplotlib

Education

- **Internationl Institute Of Information Technology** Hyderabad, Telangana
  Ph.D in Computer Science; SGPA: (8.65/10.0) *July 2012 – Present*
  - Dissertation Topic: Scalable High Dimensional Data Clustering.
  - Recipient of TCS Research Fellowship.
  - Coursework in Machine Learning, Data Mining and Warehouse, Natural Language Processing, Information Retrieval, Advance Problem Solving.
- **Jalpaiguri Government Engineering College** Jalpaiguri, West Bengal
  Bachelor of Engineering in Computer Science; SGPA: (8.55/10.0) *July 2006 – July. 2010*

Experience

- **Data Science & Analytics Center** International Institute of Information Technology, Hyderabad
  Research Assistant *July 2012 - Present*
  - **Anamoly detection framework** : Leading the team working on anamoly detection on various lab parameter values of patients useful in clinical drug trial. Funded by Novartis, Hyderabad
  - **Gene Association Extraction** : Led the team to extract cancer-gene-treatment associations from the medical text. Funded by Innovius Health.
  - **Medical Document Analysis – phase II**: Led the team working on end-to-end module for text analysis, visualization of medical documents and cohort of patients. Funded by Hitachi R & D, Bangalore
  - **Medical Document Analysis – phase I**: Led the team working on extraction and classification of named entities and relationship from medical documents. Funded by Hitachi R & D, Bangalore
  - **Twitter Sentiment Analysis**: Implemented the sentiment analysis module over twitter stream which included developing the machine learning and data visualization module. Funded by Hitachi R & D, Singapore
- **International Institute of Information Technology** Hyderabad, Telangana
  Teaching Assistant *July 2014 - December 2016*
  Natural Language Processing, Data Warehouse & Data Mining, Data Structures
- **Tata Consultancy Services** Kolkata, West Bengal
  Software Engineer *September 2010 – June 2012*
  - **Haulier and Supplier Management**: Created a haulier and supplier management system using dotnet framework, for a retail client. Participated in all the phases of the project development including software requirement gathering, prototype building and designing UI and then development. Worked on both frontend and backend components of the system and later on documentation.

Projects

- **Anomaly detection in patient lab values**: Feedback based anomaly detection framework for identifying and visualizing abnormal lab values of patients.
  Tools and technology : python, scikit learn, pytorch, matplotlib.
- **Deep Clustering and Outlier Detection**: A semi-supervised deep clustering framework for simultaneous clustering and outlier/noise detection in high dimensional data.
  Tools and Technologies used : py-torch
- **Patient Cohort Detection and Visualisation**: A weak supervised metric learning system for patient similarity detection from discharge summaries which is further used for identifying patient cohorts using clustering and visualisation.
  Tools and Technologies used : keras, D3.js, UMLS.
- **Overcoming Data Sparsity in Neural Machine Translation**: Effective Neural Machine Translation system for data sparse Indian language pairs using weak supervised learning and features.
  Tools and Technologies used : Open NMT, Anusaaraka, python
- **GeneAssociationExtractor**: Given a query, this toolkit first identifies relevant research papers. Further, it identifies putative disease-gene, drug-gene associations
  Tools and Technologies used : keras, svmlight, crf++, python.
- **MedExtract**: A toolkit for extraction of named entities like medication, disease, treatment etc and their relationship like medication-disease, treatment-disease and its classification into classes like treatment improves condition, treatment deteriorates condition and drug-adverse reaction identification from medical document.
  Tools and Technologies used : keras, libsvm, crf++, python.
- **ChatMe**: A personalized chatbot application using seq2seq architecture.
  Tools and Technologies used : Open NMT
- **Author paper Identification (Kaggle)**: Determine whether an author has written a given paper.
  Tools and Technologies used : LibSVM, Stanford NLP, JAava
- **Twitter Sentiment Analysis**: Sentiment Analysis and visualisation toolkit for twitter data.
  Tools and Technologies used : Java, Lucene, Stanford NLP, libsvm
- **Finding Most Influential Entities on Web**: A system to find and rank most influential people among a group of Baidu users differentiating between fake and real users.
  Tools and Technologies used : Graphchi, Java

Publication

- Mihir Shekhar, Lini Thomas, Kamalakar Karlapalem, High Dimensional Clustering: A Strongly Connected Component Clustering Solution (SCCC) , ICDM 2018 (Under Review)
- Ruchit Agarwal, Mihir Shekhar, Dipti Mishra Sharma, Three-phase training to address data sparsity in Neural Machine Translation 2017, ICON(oral)
- Ruchit Agarwal, Mihir Shekhar, Dipti Mishra Sharma, Integrating knowledge encoded by linguistic phenomena of Indian Languages with Neural Machine Translation , 2017, MIKE(oral)
- Mihir Shekhar, Raghavendra Ch., Lini Thomas, Sunil Mandhan, Kamalakar Karlapalem, Identifying Medical Terms Related to Specific Diseases . 2015 Biological Data Mining and its Applications in Healthcare (ICDM 2015 workshop).
- Mihir Shekhar, K Santosh, Romil Bansal, Vasudeva Varma, Author Profiling: Predicting Age and Gender from Blogs, Notebook for PAN at CLEF 2013, Valencia, Spain.

Accomplishment

- Obtained TCS Research Scholarship
- Obtained second position in Author Profiling Task in CLEF 2013
- TCS ILP top performer 2010
- Top 20 percentile in KDD Cup 2013, Author Paper Identification Challenge

Reference

Prof. Kamalakar Karlapalem
https://faculty.iiit.ac.in/ kamal/
Head of Department, Data Science and Analytics Centre.
International Institute of Information Technology, Hyderabad