Soumyadeep Roy

Complex Networks Research Lab, Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur - 721302, West Bengal, India
soumyadeep.roy9@iitkgp.ac.in • https://roysoumya.github.io/ • github.com/roysoumya

EDUCATION

Indian Institute of Technology Kharagpur, West Bengal, India

- **Ph.D. Candidate in Computer Science and Engineering** Jan 2020 – Present
  - Thesis: Incorporating Domain Knowledge in Medical NLP Applications
  - Advisor: Prof. Niloy Ganguly and Prof. Shamik Sural
  - Focus areas: Artificial Intelligence in Medicine, Natural Language Processing, Medical Text Summarization
  - Supported by Institute Ph.D. Fellowship.

- **M.S. (Research) in Computer Science and Engineering** Jul 2017 – Nov 2019
  - Thesis: Computational Approaches for Online Reputation Monitoring [Modeling, Analysis and Recommendation]
  - Advisor: Prof. Niloy Ganguly and Prof. Shamik Sural
  - Cumulative GPA: 9.21 / 10.00 ; Thesis

Kalyani Government Engineering College, Kalyani, West Bengal, India

- **B.Tech in Computer Science and Engineering** Jul 2013 – Jun 2017
  - Thesis : Understanding Email Interactivity and Predicting User Response to Email
  - Advisor: Dr. Kousik Dasgupta and Mr. Binay Gupta
  - Cumulative GPA: 8.61 / 10.00

RESEARCH EXPERIENCE

L3S Research Center, Leibniz University Hannover, Germany

- Worked as a Research Associate in *International Leibniz Future Laboratory for Artificial Intelligence* with Prof. Wolfgang Nejdl. Here, I worked on developing efficient pretraining strategies for the gene foundational models, which led to an ECAI 2023 publication. I also worked on the medical use case of Parkinson's disease in collaboration with Hannover Medical School. Germany, where we developed a decision tree-based patient subtyping method for Parkinson's Disease. Additionally, I was the point person who organized the meetings and coordinated the reports for this use case. Jan 2021 – Jun 2023

Adobe Systems India Pvt. Ltd., Bangalore, India

- Worked as a Summer Research Intern with Mr. Anandhavelu Natarajan and Dr. Niyati Chhaya of the Adobe Big Data Experience Lab Bangalore on the topic "Predicting Brand Personality from Web Content". Here, we developed a deep learning-based brand personality classification model and constructed a labeled brand personality detection dataset using company webpages by annotating with the Amazon Mechanical Turk platform. May 2018 – Jun 2018

Indian Institute of Technology Kharagpur, West Bengal, India

- Selected for the Indian Academy of Sciences Summer Research Fellowship Programme to work with Prof. Sudeshna Sarkar of the Computer Science and Engineering Department on the project topic "Anomaly Detection and Summarization of various Extreme Weather Events over Indian sub-continent". Here, we formulated the extreme weather event detection problem as an anomaly detection problem over spatiotemporal weather datasets and used cSTAG, a graph-based algorithm for identifying patterns of correlated change to detect extreme weather events. Jun 2016 – Jul 2016
  - Technical Report, Code

SELECTED PUBLICATIONS

CONFERENCES

[ECAI 2024] S. Roy, S. Sural, N. Ganguly, "Unlocking Efficiency: Adaptive Masking for Gene Transformer Models," in the *27th European Conference on Artificial Intelligence*, 8 pages, Santiago de Compostela, Spain, Oct 2024, arXiv code official accepted papers list (to appear)

[IJCAI 2024] G. Balde*, S. Roy*, M. Mondal, N. Ganguly, "MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization," in the *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024 Main Track)*, 9 pages, Jeju, South Korea, Aug 2024, doi: 10.24963/ijcai.2024/683 arXiv code (* for equal contribution)

[SIGIR 2024]	S. Roy, A. Khatua, F. Ghoochani, U. Hadler, W. Nejdl, N. Ganguly, "Beyond Accuracy: Investigating Error Types in GPT-4 Responses to USMLE Questions," in the *47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024 Resource and Reproducibility Track)*, pages: Pages 1073 - 1082, Washington D.C., USA, Jul 2024, doi: 10.1145/3626772.3657882 arXiv code
[ECAI 2023]	S. Roy, J. Wallat, S. Sundaram, W. Nejdl and N. Ganguly, "GeneMask: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning," in the *26th European Conference on Artificial Intelligence (ECAI 2023)*, 8 pages, Krakow, Poland, Oct 2023, doi: 10.3233/FAIA230492 arXiv code slides **Received 3rd prize in poster presentation of IndoML 2023 at IIT Bombay**
[ICDH 2023]	S. Roy, N. Ganguly, S. Sural and K. Rudra, "Interpretable Clinical Trial Search using Pubmed Citation Network," in the *2023 IEEE International Conference on Digital Health (ICDH)* , pp. 328 - 338, Chicago, United States, Jul 2023, doi: 10.1109/ICDH60066.2023.00056 pdf code slides **Candidate for Best Student Paper**
[CIKM 2021]	S. Roy, S. Chakraborty, A. Mandal, P. Sharma, G. Balde, A. Natarajan, M. Khosla, S. Sural and N. Ganguly, "Developing Knowledge-Aware Neural Models for Medical Forum Question Classification," in the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), 1-5 November 2021, Online, 4 pages doi arXiv code slides video
[WebSci 2019]	S. Roy, N. Ganguly, S. Sural, N. Chhaya and A. Natarajan, "Understanding Brand Consistency from Web Content," in *Proceedings of the 10th ACM Conference on Web Science (WebSci 2019)*, pp. 245 - 253, Boston, MA, USA, Jul 2019. doi pdf, data, slides

JOURNALS

[TWEB 2021]	S. Roy, S. Sural, N. Chhaya, A. Natarajan and N. Ganguly, "An Integrated Approach for Improving Brand Consistency of Web Content: Modeling, Analysis and Recommendation," in ACM Transactions on the Web, volume 15, article 2, Article 9 (May 2021), 25 pages. doi arXiv code and data
[Work-in-progress]	S. Roy, S. Mücke, M. Marschollek, H. Frieling, N. Ganguly and D. Wolff, "Decision Tree-based Approach to Robust Parkinson's Disease Subtyping using Clinical Data of the Michael J. Fox Foundation LRRK2 Cross-sectional Study," ***Work-in-progress*** *at Communications Medicine, collaboration with Medical School Hannover, Germany*

PROFESSIONAL ACTIVITIES (TALKS/ TRAVEL GRANTS/ REVIEWING)	▪ Received the Google Research Travel Grant for our IJCAI 2024 paper ▪ Reviewed for Core A* conferences: EMNLP 2022, AAAI 2023, ACL 2023, EMNLP 2023, AAAI 2024, AAAI 2025. Also reviewed for EACL 2023, ECML-PKDD 2020 (Core A), IEEE Transactions on Cognitive and Developmental Systems (Journal, Impact Factor: 5), Elsevier Artificial Intelligence In Medicine (Journal, Impact Factor: 6.1) ▪ Received the Microsoft Research Travel Grant for attending ECAI 2023 at Krakow, Poland, from Sept. 30 to Oct. 5, 2023 ▪ Received the SIGIR Student Travel Award for attending CIKM 2021 (virtual) ▪ Performed the role of *Subject Matter Expert* for the "Data Science and Machine Learning Track" of Pycon India 2020 and part of "Mentorship Team" of Pycon India 2021. ▪ Presented the ACM WebSci 2019 paper titled *Understanding Brand Consistency from Web Content* at the "Out-of-India Exhibition" track of India HCI 2019 held at the Indian School of Business, Hyderabad, India, on November 1 to 3, 2019 ▪ Received IIT Kharagpur's *Institute Travel Grant* and CNeRG Travel Grant for attending ACM WebSci 2019 at Northeastern University, Boston, MA, United States from July 1 - 6, 2019. ▪ Received *Best Graduate Forum Presentation Award* for the paper "Automated EBM-oriented Summarization of Active or Recruiting Clinical Trials," in the *10th International Conference on COMmunication Systems & NETworkS (COMSNETS 2018)*, 2 pages, Bangalore, India
TEACHING (TA-SHIP)	▪ **IIT Kharagpur:** Programming and Data Structure Lab (Spring 2018), Department Website Team (Autumn 2018 to Spring 2020), Information Retrieval (Autumn 2020), Computing Lab 1 (Autumn 2023) ▪ **Leibniz University Hannover, Germany:** Natural Language Processing (Winter 2021), Foundations of Information Retrieval (Summer 2022, 2023)

MENTORING	▪ **3 M.Tech Thesis:** (i) *Multidimensional Retrieval and Ranking Model for Clinical Trials Search* by Nikhil Agrawal (2019), (ii) *Query-Focused Summarization in Medical Domain: A Comparative Study* by Gunjan Balde (2020), and (iii) *Understanding Medical Knowledge Graph for NLP Applications* by Sourav Saha (2021)
	▪ **1 B.Tech Thesis** titled *Improving Question Generation of Questions Answering Based Summary Evaluation Metric* by Sudip Chakraborty (2022)
REFERENCES	▪ **Prof. Niloy Ganguly**, Professor of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal, India. http://www.facweb.iitkgp.ac.in/ niloy/ Email id: niloy@cse.iitkgp.ac.in
	▪ **Prof. Wolfgang Nejdl**, Professor of Computer Science, Leibniz University Hannover, L3S Research Center, Hannover, Germany. https://www.kbs.uni-hannover.de/ nejdl/ Email id: nejdl@l3s.de
	▪ **Dr. Mainack Mondal**, Assistant Professor of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal, India. https://cse.iitkgp.ac.in/ mainack/ Email id: mainack@cse.iitkgp.ac.in
	▪ **Prof. Michael Marschollek**, Professor of Medical Informatics, Hannover Medical School, Technische Universität Braunschweig, Germany. https://plri.de/team/michael_marschollek Email id: marschollek.michael@mh-hannover.de