

Soumyadeep Roy

Center for Biomedical Informatics Research, Stanford University, California, USA
soroy@stanford.edu • <https://roysoumya.github.io/> • github.com/roysoumya • Stanford Profile Page

PHD SUMMARY

I am a postdoctoral scholar at the Center for Biomedical Informatics Research of Stanford University, advised by Prof. Tina Hernandez-Boussard. My primary area of research is natural language processing, with expertise in medical and healthcare applications. My research areas of interest are Foundation Models for Medicine, Generative AI, Text Summarization, and Efficient Pretraining.

I hold a PhD in Computer Science and Engineering from the Indian Institute of Technology Kharagpur, where I worked with Prof. Niloy Ganguly and Prof. Shamik Sural. Here, I was part of the Complex Networks Research Group (CNeRG). My PhD thesis is titled “Domain Adaptation for Medical Language Understanding”, where I developed novel domain adaptation techniques to effectively and efficiently adapt open-domain AI models to the medical domain.

EDUCATION

Indian Institute of Technology Kharagpur, West Bengal, India

- PhD in Computer Science and Engineering Jan 2020 – Jul 2025
 - Thesis: Domain Adaptation for Medical Language Understanding
 - Advisor: Prof. Niloy Ganguly and Prof. Shamik Sural
 - Joint MS-PhD program
- MS 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

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

WORK EXPERIENCE

Center for Biomedical Informatics Research, Stanford University, California, USA Sep 2025 – present

- Postdoctoral Scholar (Biomedical Informatics) advised by Prof. Tina Hernandez-Boussard

Wipro GE Healthcare Pvt. Ltd., Bangalore, India

Nov 2024 – May 2025

- PhD Research Intern at GE Healthcare Innovation and Technology Center. I worked on two projects:
 - (i) Deep learning-based representation learning of sensor log data from medical imaging equipment, and
 - (ii) Generative AI-based dashboards of oncology patients to help clinicians easily navigate patient notes.

L3S Research Center, Leibniz University Hannover, Germany

Jan 2021 – Jun 2023

- Research Associate (*Wissenschaftlicher Mitarbeiter* in German) with Prof. Wolfgang Nejdl of L3S Research Center. I worked on the medical use case of Parkinson’s disease in collaboration with Hannover Medical School, Germany, where we analyzed the clinical data of Parkinson’s Disease patients and developed a decision tree-based patient subtyping method for Parkinson’s Disease.

Adobe Systems India Pvt. Ltd., Bangalore, India

May 2018 – Jun 2018

- Research Intern with Mr. Anandhavelu Natarajan and Dr. Niyati Chhaya of the Adobe Big Data Experience Lab in Bangalore, on the topic “Predicting Brand Personality from Web Content”. Here, I developed a deep learning-based brand personality classification model and led the data annotation setup on Amazon Mechanical Turk.

Indian Institute of Technology Kharagpur, West Bengal, India

Jun 2016 – Jul 2016

- Summer Research Intern with Prof. Sudeshna Sarkar on the project topic “Anomaly Detection and Summarization of various Extreme Weather Events over Indian sub-continent”. I formulated the extreme weather event detection problem as an anomaly detection problem over a spatio-temporal graph and implemented a graph-based clustering algorithm to identify extreme events based on the detected clusters.

SELECTED PUBLICATIONS

CONFERENCES

- [ECAI 2025] [T. Rohr](#), S. Roy, F. Mohamad, J. Papaioannou, W. Nejdl, F. Gers, A. Loser, “Where does it hurt?” - Dataset and Study on Physician Intent Trajectories in Doctor Patient Dialogues,” in the *28th European Conference on Artificial Intelligence*, 8 pages, Bologna, Italy, Oct 2025 [arXiv code](#) (to appear)
- [ACL 2025] G. Balde, [S. Roy](#), M. Mondal, N. Ganguly, “Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings,” in the *63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025 Findings Full Paper*, Vienna, Austria Jul 2025 [arXiv code](#) [doi](#)
- [WSDM 2025] S. Roy, S. Sundaram, D. Wolff, N. Ganguly, “Building Trustworthy AI Models for Medicine: From Theory to Applications,” in the *18th ACM International Conference on Web Search and Data Mining*, Tutorial paper, pages 1012 - 1015, Hannover, Germany, Mar 2025 [doi](#)
- [WSDM 2025] N. Steiner, Z. Li, O. Vosoughi, J. Schrader, [S. Roy](#), W. Nejdl, M. Tang, “A Systematic Evaluation of Single-Cell Foundation Models on Cell-Type Classification Task,” in the *18th ACM International Conference on Web Search and Data Mining, WSDM Day*, pages 1112 - 1113, Hannover, Germany, Mar 2025, [doi](#)
- [EMNLP 2024] G. Balde*, [S. Roy*](#), M. Mondal, N. Ganguly, “Adaptive BPE Tokenization for Enhanced Vocabulary Adaptation in Finetuning Pretrained Language Models,” in the *The 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP Findings Short Paper*, Miami, Florida, Nov 2024, [doi](#) [arXiv code](#) (* for equal contribution)
- [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, [doi](#) [arXiv code](#)
- [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](#) [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](#) [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](#) [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](#) [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](#)

[Under Review] 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,” *Under review at Frontiers in Artificial Intelligence Journal; Collaboration with Hannover Medical School, Germany*

**PROFESSIONAL
ACTIVITIES
(TALKS/ TRAVEL
GRANTS/
REVIEWING)**

- Invited to present our ECAI 2023 and 2024 work on “Improving Training Efficiency of DNA Foundation Models” at the 1st Symposium on Intersections: Statistical Genomics and Complex/Chronic Disease Biology, organised by BRICS - National Institute of Biomedical Genomics, Kalyani, West Bengal
- Delivered an invited talk on “Foundation Models for Medical Text” on December 19, 2024, as part of the course “AI Foundation Models in Biomedicine” of Leibniz University Hannover, Germany
- Received the ACM-IARCS Conference Travel Grant for presenting the ECAI 2024 paper in Santiago de Compostela, Spain
- Reviewed for **Core A* conferences**: EMNLP 2022, AAAI 2023, ACL 2023, EMNLP 2023, AAAI 2024, AAAI 2025, AAAI 2026 ACL Rolling Review February, May, and July 2025 cycle.
- Reviewed for **other conferences**: COLING 2025, ML4H 2024, EACL 2023, ECML-PKDD 2020 (Core A).
- Reviewed for **journals** - Knowledge and Information Systems, Frontiers in Genetics, IEEE Transactions on Cognitive and Developmental Systems, Elsevier Artificial Intelligence in Medicine
- 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.
- 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

Can be provided on request.