Punyajoy Saha

PhD student, Computer Science and Engineering

Indian Institute of Technology, Kharagpur, West Bengal, India - 721302

punyajoysaha1998@gmail.com | https://punyajoy.github.io | punyajoy | punyajoy-saha | @punyajoysaha

Research Summary

My research lie in the intersection of computational social science and natural language processing. Currently, I am working in developing better human-in-the-loop mitigation strategies for hate speech and other forms of harmful speech like fear speech. I have experience in curating high quality social datasets and building natural language processing solutions. I am deeply interested in AI for social good.

Education

Indian Institute of Technology, Kharagpur *Kharagpur, West Bengal, India*

PhD Computer Science and Engineering *Jan 2020 - present*

- **CGPA**: 9.13 ; **Advisor**: Dr. Animesh Mukherjee ; Member of CNeRG labs
- PhD Thesis: Enabling moderation of harmful content in online social media platforms

Indian Institute of Engineering Science and Technology, Shibpur *Shibpur, Howrah, India*

B.Tech Undergraduate Degree *2015 - 2019*

- **CGPA** of 8.93 ; **Advisor**: Dr. Jaya Sil
- Computer Vision Head at Robodarshan, the robotics club of IIEST Shibpur (2017-18)
- B.Tech thesis: Application of Generative Adversarial Networks for Wallpaper Generation on Digital Devices (Link)

Research Experiences

Georgia Institute of Technology *Atlanta, United States of America*

Research Visitor *Jun 2023 - Sep 2023*

- Was invited in the lab of Prof. Srijan Kumar's Computational Data Science Lab for the Web and Social Media
- Worked on understanding the impact of counterspeech which is a novel to respond to hate speech
- Built a pipeline for identifying counter speech in the wild using language modelling techniques

University of Hamburg *Hamburg, Germany*

Research Visitor *May 2022 - Aug 2023*

- Worked with Dr. Abhik Jana and Dr. Chris Biemann in the Language Technology Group
- The first project was building a human-in-the-loop method for counterspeech generation. This pipeline is currently being used for developing a diverse counterspeech dataset.
- The second project involved understanding the intrinsic properties of language models in generation of counterspeech. We are also trying prompt based techniques to evaluate type specific counter speech generation.

University of Hamburg *Hamburg, Germany*

Research Visitor *Aug 2019 - Sep 2019*

- Worked with Dr. Gregor Wiedemann and Dr. Chris Biemann in the Language Technology Group
- Developed the backend for Forum 4.0 project on providing a text analytics software for journalists

Publications

Roy, S., Harshavardhan, A., Mukherjee, A. and **Saha, P**., (2023). Probing LLMs for hate speech detection: strengths and vulnerabilities. arXiv preprint arXiv:2310.12860. (accepted in EMNLP 2023)

Saha, P., Sheth, D., Kedia, K., Mathew, B., & Mukherjee, A. (2023). Rationale-Guided Few-Shot Classification to Detect Abusive Language. In Proceedings of the 26th European Conference on Artificial Intelligence (pp. 2041-2048)

Das, M., Raj, R., **Saha, P**., Mathew, B., Gupta, M., & Mukherjee, A. (2023). HateMM: A Multi-Modal Dataset for Hate Video Classification. Proceedings of the International AAAI Conference on Web and Social Media, 17(1), 1014-1023.

Aggarwal, P., Chawla, P., Das, M., **Saha, P**., Mathew, B., Zesch, T. & Mukherjee, A., (2023, April). HateProof: Are Hateful Meme Detection Systems really Robust?. In Proceedings of the ACM Web Conference 2023 (pp. 3734-3743).

Gupta, V., Roychowdhury, S., Das, M., Banerjee, S., **Saha, P**., Mathew, B. and Mukherjee, A., (2022). Multilingual Abusive Comment Detection at Scale for Indic Languages. Advances in Neural Information Processing Systems, 35, pp.26176-26191.

Saha, P., Garimella K., Kalyana N.K., Pandeya S.K., Meher P., Mathew B., & Mukherjee A. (2022). On the rise of fear speech in online social media. In Proceedings of the National Academy of Sciences of the United States of America (PNAS).

Das, M., Dash, A., Jaiswal, S., Mathew, B., **Saha, P**. and Mukerjee, A., 2022. Platform Governance: Past, Present and Future. GetMobile: Mobile Computing and Communications, 26(1), pp.14-20.

Saha, P., Singh, K., Kumar, A., Mathew, B., & Mukherjee, A. (2022). CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (pp.5157-5163)

Das, M., Banerjee, S., **Saha, P**. and Mukherjee, A., 2022, November. Hate Speech and Offensive Language Detection in Bengali. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (pp. 286-296).

Das, M., **Saha, P**., Dutt, R., Goyal, P., Mukherjee, A., & Mathew, B. (2021, August). You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (pp. 79-89).

Mathew, B., **Saha, P**., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 14867-14875).

Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021, April). "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In Proceedings of the Web Conference 2021 (pp. 1110-1121).

Aluru, S. S., Mathew, B., **Saha, P**., & Mukherjee, A. (2020, September). A Deep Dive into Multilingual Hate Speech Classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 423-439). Springer, Cham.

Mathew, B., **Saha, P**., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., & Mukherjee, A. (2019, July). Thou shalt not hate: Countering online hate speech. In Proceedings of the international AAAI conference on web and social media (Vol. 13, pp. 369-380).

Awards, Honours & Fellowships

2023	**Invited to Heidelberg Laureate Forum**, HLF	
	Grant accepted "Deep neural multilingual models to combat online hate content", SPARC	
2022	**Invited to Google research days**, Google	
2021	**Ted Nelson Newcomer Award (Link)**, HyperText	*$ 1000*
	Winner of Abusive and Threatening Language Detection Task (Link), FIRE	*$ 280*
	Winner of Offensive language shared task, DravidianLangTech	
2020	**World Rank 11 in Facebook Hatememe detection competition (Link)**, Meta AI	
	PhD Scholarship, PMRF	*Rs 70,000 pm*
2019	**1st in Offensive speech detection**, HASOC	
2018	**1st in Misogynistic text classification**, AMI-EVALITA	

Tutorials and Invited talks

Invited talk "Echoes of Fear: Unraveling the Presence of Fear Speech in Social Media Platforms." in CyberVSR'23, RIT (Link)

Tutorial on "Hate speech: Detection, Mitigation and Beyond" accepted at WSDM 2023, Singapore. (Link)

Tutorial on "Hate speech: Detection, Mitigation and Beyond" accepted at AAAI 2022 (Link).

Presented the "Fear speech in Whatsapp Groups" in Stanford Internet Observatory's End-to-end encryption Workshop

Invited talk in Understanding and Automating Counterspeech workshop, 2021 organized by CRASSH

Tutorial on "Hate speech: Detection, Mitigation and Beyond" accepted at ICWSM 2021 (Link).

Invited talk on the topic of "Generative Adversarial Network" IIEST, Shibpur, 2020 (Online).