Letter of Interest
Predoctoral Researcher

Aatman Vaidya
CV | Website

I am interested in using large scale computational methods with human centered studies to understand societal phenomena. I seek to understand how new technologies and practices shape the web, specifically social media, examine their impact on human lives, and the broader societal consequences of this interplay. These interests stem from having research experience in civic tech and building tools to respond to online abuse and harms in India. My research interests broadly lie under computational social science, primarily in natural language processing and network science.

Research Experience I wanted to better understand how hate spreads on a social network, hence I decided to channel my undergraduate thesis towards proposing a model that maps the spread of toxicity on X (formerly Twitter). An in-depth empirical analysis led us to find that existing work like spread activation models and epidemic models like susceptible-infected-recovered (SIR) inadequately capture the nuances and spread of hateful content. We proposed a model that is based on user behaviour and captures two important factors: a) toxicity exists as a spectrum, i.e. hate is not binary, and b) toxicity is not conserved in a network. This led to a first-authored work published [1]. During this project, I understood how pervasively hateful content spreads through a network and reaches users. I learnt more about the process of research itself and, most importantly, how to solve larger problems by breaking them down into actionable steps.

Interventions for Online Harms Having worked on modeling harmful content, building strategies to mitigate it was a natural progression to my research interests. I currently work at Tattle Civic Tech[1] where I help build citizen-centric tools and datasets to understand and respond to misinformation and harmful content online in India.

I work on a browser plugin, Uli that redacts abusive content and collectively helps push back against online gender-based violence (oGBV). I programmed a production scale feature where users at the receiving end of online abuse could crowdsource metadata[2] associated with abusive words, which is crucial for better contextualization of harmful content for NLP models. This dataset will be useful to trust and safety teams of social media companies to improve content moderation for Indian Languages. I performed a literature survey, data analysis and handled all logistics, for a research project on an expertly annotated dataset on online gender-based violence[3] [2]. As the dataset was annotated by activists and researchers who have experienced oGBV, I gathered important insights on handling such complex subjective annotations, such as seeking qualitative insights to understand annotator disagreements better.

I am currently working on a configurable engine called Feluda for analyzing multilingual and multimodal content, where I am building methods to analyze audio and video data. I build data pipelines that help detect similar media items, extract key information and claims, and help remove spam. This work is being used as a part of a broader Whatsapp Helpline to respond to misinformation and deepfakes around elections in India [3]. Addressing the need of fact-checkers, as a part of Feluda, I created workflows to cluster large amounts of videos into different thematic social media labels like political speeches, news interviews, devotional content, short-form

[1] https://tattle.co.in/
[2] Metadata such as was this word used casually? is it appropriated? context it was used in etc
[3] This work won an Outstanding Paper award at the Workshop of Online Harms and Abuse, NAACL '24.

reels, spam etc. I did a thorough evaluation of different clustering algorithms, vision language models (VLMs) and created an extensive custom video dataset in an Indian context to carry out experiments.

Working on these projects at Tattle Civic Tech, I was exposed to literature on bias, evaluation, and transparency [4], especially how hate speech and misinformation adversely affect marginalised communities in India. I learnt how to combine feminist principles with technical methods to build responsible AI and how tech-enabled interventions can help mitigate social problems [5]. I learnt an ethics-informed approach towards building AI tools and systems. I also developed important engineering skills in working with large-scale deployment and writing production-level code. This has greatly shaped my values as a person, and building AI for social good will be central to my long-term career goals.

Future Research Interest Drawing from my research experiences, I want to understand how new technologies and practices shape the web, specifically social media, and their impact on human lives. I am interested in taking a **human-centred** and **community-led** approach towards how AI can be made inclusive, safer and impactful to society at large. A key focus of my future research is to improve online safety and make online communities safer. This is informed from working with social activists, researchers and organizations who are working at the grassroots of these problems. Social media algorithms often amplify harmful content and fail to account for social-context, this motivates me to work towards designing better algorithms and systems, ones that are more socially aware, contextually grounded and provide transparency.

Why this Predoctoral Researcher role? My recent work on studying online platforms has pushed me to think about how online platforms shape democratic/political discourse and influence user behavior, and how it can lead to real world harm. This predoctoral position is an ideal opportunity to further that inquiry. **Prof Ashiwn's** recent work on framing news coverage, modelling personal narratives in political discussions is essential in pushing this area of inquiry forward. I want to work with them on also understanding similar questions like, how do algorithmic amplification and influencer networks shape political discourse and agenda-setting on social media platforms? How do political narratives emerge, evolve cross platform?

In the future, I see myself working on problems at the intersection of AI, society and language. My short term goal is to do a PhD in computational social science, my long term goal is to make a research career in academia or industry, hence a Predoctoral Researcher is the natural next step towards it.

References

[1] **Aatman Vaidya**, Seema Nagar, and Amit A. Nanavati. *"Analysing the Spread of Toxicity on Twitter."* In Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD **CODS** and 29th **COMAD**), pp. 118-126. 2024.
[2] Arora, Arnav, Maha Jinadoss ... **Aatman Vaidya**, Tarunima Prabhakar et at, "*The Uli Dataset: An Exercise inExperience Led Annotation of oGBV*", Workshop on Online Abuse and Harms, NAACL 2024
[3] Media Coverage: Nieman Lab | The Hindu
[4] **Aatman Vaidya**, Tarunima Prabhakar, Denny George, Swair Shah, "*Analysis of Indic Language Capabilities in LLMs*", arXiv:2501.13912
[5] Tarunima Prabhakar, Srravya Chandhiramowuli, Vamsi Krishna Pothuru, Saumya Gupta, Mansi Gupta, Kaustubha Kalidindi, **Aatman Vaidya**, "*AI Safety Benchmark Dataset for Hate and Sex-Related Crimes in Hindi*", Part of the AILuminate v1.1 benchmark suite by MLCommons.