My research interests primarily lie in Natural Language Processing (NLP) and Social Network Analysis (SNA), particularly in building **multilingual** and **multimodal** models for low-resource languages. In addition to core computer science courses, I also took courses in anthropology, Indian philosophy and developmental sociology. These courses prompted me to ask a fundamental question, "How do technology and human interactions shape and influence each other? What are the broader societal consequences of this interplay?"

Having grown up in the digital age, I sought to answer this question, particularly in the context of social media. State-of-the-art NLP models show great results on benchmarks, but they often struggle with socio-cultural understanding, do not work well for low-resource languages and may perpetuate harmful societal biases and stereotypes. This is witnessed in the social media landscape, where content moderation for Indian languages is very poor.

Modelling Harmful Content. Under the guidance of Prof. Amit Nanavati and Dr Seema Nagar (IBM Research India), I decided to channel my undergraduate thesis towards proposing a novel model that maps the spread of toxicity[1] on X (formerly Twitter). An in-depth empirical analysis led us to find that existing work like Spread Activation models (SPA) and epidemic models like Susceptible-Infected-Recovered (SIR) inadequately capture the nuances and spread of hateful content. Our model 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 at **CODS-COMAD 2024** [1].

We extended this work by studying toxicity as a time series and forecasted it using transformer-based models and graph CNNs. While network structure matters, we found that explicit connections may not always imply influence. This study led to a first-authored work published at **IEEE CogMI 2023** [2]. During this project, I realised 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.

Building Systems To Combat Hate and Misinformation. Having worked on modelling harmful content, mitigating it was a natural progression for my research interests. At Tattle Civic Tech, I build tools and datasets to understand and respond to inaccurate and harmful content. I contribute to an open-source browser plugin called "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 associated with abusive words, which is crucial for better contextualisation of harmful content for NLP models. This open-source dataset will be useful to trust and safety teams of social media companies to improve content moderation for Indian Languages. I also performed a literature survey and data analysis for a research project on a dataset about online gender-based abuse **[3]**. 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.

To extend this, I led a shared task conducted at **ICON 2023** [4] to detect oGBV and saw that there was a lack of sufficient data points for models to generalise well. This could be improved by augmenting the dataset with more instances. I am currently working on a configurable engine called "Feluda" for analysing multilingual and multimodal content, where I am building methods to analyse audio and video data. I performed an in-depth analysis of state-of-the-art optical character recognition (OCR) models for Indian languages and provided insights

[1] Toxicity is defined as a "*rude, disrespectful, unreasonable comment that is likely to make someone leave a conversation*". I use toxicity and hate interchangeably.

into improving them. My analysis of these models and work on Feluda will aid fact-checkers in responding to misinformation.

Working on these projects at Tattle Civic Tech, I was exposed to literature on fairness and transparency, 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 the interplay of tech and policy can help mitigate social problems. I 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 Interests Drawing from my research experiences, 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. I want to build holistic multilingual NLP models that are socially aware. Language data in low-resource languages is often of poor quality; hence, instead of creating general benchmark datasets, I want to take an intersectional approach and create datasets to address specific harms such as oGBV, communal aggression, etc. These datasets could fill the resourcedness gap and create multilingual LLMs that show strong generalisation ability across different languages [5]. I want to leverage a hybrid solution of LLM-based evaluators and native language speakers to scale up multilingual evaluation [6]. Humans don't just comprehend text on their own but use their life experiences and context from multimodal sources. I want to build models that represent a close integration between language and our world, improving a machine's language understanding capability.

In the future, I see myself working on problems at the intersection of AI, society and language. In the long term, I wish to pursue a PhD in social computing.

[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] **Aatman Vaidya**, Seema Nagar, and Amit A. Nanavati. "Forecasting the Spread of Toxicity on Twitter." In Proceedings of the 5th IEEE International Conference on Cognitive Machine Intelligence (**IEEECogMI**).

[3] Arora, Arnav, Maha Jinadoss .. **Aatman Vaidya**, Tarunima Prabhakar et at, "The Uli Dataset: An Exercise inExperience Led Annotation of oGBV"arXiv preprint arXiv:2311.09086

[4] **Aatman Vaidya**, Arnav Arora, Aditya Joshi, and Tarunima Prabhakar. "Overview of the 2023 ICON Shared Task on Gendered Abuse Detection in Indic Languages." 20th International Conference on Natural Language Processing (**ICON**).

[5] Nicholas, G., & Bhatia, A. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis. arXiv preprint arXiv:2306.07377.

[6] Hada, R., Gumma, V., de Wynter, A., Diddee, H., Ahmed, M., Choudhury, M., ... & Sitaram, S. (2023). Are large language model-based evaluators the solution to scaling up multilingual evaluation?. arXiv preprint arXiv:2309.07462.