Aatman **Vaidya**

SDE | Tattle Civic Tech

aatmanvaidya.github.io | aatmanvaidya@gmail.com | aatmanvaidya | aatmanvaidya | Google Scholar

Education

Jul 2023 | **Ahmedabad University** — **Ahmedabad, India**
Aug 2019 | B.Tech in Computer Science & Engineering (CSE)
Thesis: Modelling the spread of Hate Speech on Twitter | *Advisors*: Prof. Amit Nanavati, Dr. Seema Nagar

Experience

Present / **Aug 2023** | **Tattle Civic Tech** [🌐] — **Remote**
Software Development Engineer - I and Data Engineer | Advisors: Tarunima Prabhakar, Denny George
➞ I help build citizen centric open-source tools and datasets to understand and respond to **inaccurate** and **harmful** content.
➞ I help build a browser extension Uli, that redact's slurs and abusive content, archive's problematic content, and collectively pushes back against **online gender-based violence**. Uli de-normalizes everyday violence that people of marginalized genders experience online in India. [] [Media Coverage: UNFPA Report]
➞ I lead the development of Feluda, a configurable engine for analyzing multilingual and multimodal content. Feluda helps fact-checkers and researchers in combating misinformation. [] [] [Media Coverage: UNDP]
➞ Feluda is being used as the back-end analysis engine on a national Deepfake detection **WhatsApp helpline** (Deepfakes Analysis Unit) by the Misinformation Combat Alliance for 2024 Indian Lok Sabha elections. This helpline is funded by Meta. [Media Coverage: Nieman Lab | The Hindu | Mashable | Meta]
➞ I build data pipelines to cluster large amount's of Audio and Video data into social media thematic labels like politics, humor, memes, devotional content etc. This work helps fact-checker's better understand, vizualize and analyze the content received on their helpline.
➞ Tattle engaged with a company to develop an AI safety benchmark dataset for hate and sex related crimes. As a part of this project, I also conducted a survey on evaluating Indic LLM's for natural language tasks and online harms.

May 2022 / **May 2021** | **Ahmedabad University** [🌐] — **Ahmedabad, India**
Undergraduate Research Fellow | Advisors: Prof. Shilpa Pandit, Prof. Keyur Joshi
Projects: Analysing people's psychological and physiological effects and experiences in the second wave of COVID-19, Perceivable Image Blur Quality Index

Publications

S=In Submission, C=Conference, W=Workshop, P=Preprints, J=Journal

[S.3] **Improving Content Moderation for Indian Languages using Related Tasks**
Aatman Vaidya, Dipankar Srirag, Aditya Joshi
[In Submission at WWW 2025]

[S.2] **Strategies to Mitigate Spread of Toxicity on a Social Network**
Aatman Vaidya, Harsh Bhagat, Seema Nagar, Amit Nanavati
[In preparation, ICWSM 2025]

[S.1] **Indic LLM Landscape Analysis**
Aatman Vaidya, Denny George, Swair Shah, Tarunima Prabhakar
[In preparation]

[W.1] **The Uli Dataset: An Exercise in Experience Led Annotation of oGBV**
Arnav Arora, Maha Jinadoss, Cheshta Arora, Denny George, Brindaalakshmi...Aatman Vaidya, Tarunima Prabhakar
Workshop on Online Abuse and Harms at North American Chapter of the Association for Computational Linguistics
★ Outstanding Paper Award — **[NAACL'24]**

[C.3] **Analysing the Spread of Toxicity on Twitter**
Aatman Vaidya, Seema Nagar, Amit Nanavati
ACM International Conference on Data Science and Management of Data — **[CODS-COMAD'24]**

[C.2] **Forecasting the Spread of Toxicity on Twitter**
Aatman Vaidya, Seema Nagar, Amit Nanavati
IEEE International Conference on Cognitive Machine Intelligence — **[IEEE CogMI'23]**

[C.1] **Overview of the 2023 ICON Shared Task on Gendered Abuse Detection in Indic Languages**
Aatman Vaidya, Arnav Arora, Aditya Joshi, Tarunima Prabhakar
International Conference on Natural Language Processing — **[ICON'23]**

Select Research Projects

Modelling Spread of Hate Speech Aug'22 - Present

Advisors: Dr. Amit Nanavati, Dr. Seema Nagar (IBM Research India)

- Worked on a novel model to capture the spread of hate speech on Twitter. 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. [**Full Paper@CODS-COMAD'24**]
- 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. [**Full Paper@IEEE CogMI'23**]
- An in-depth empirical analysis led us to find users change behaviour with time and is impacted by its neighborhood. Developing a model that captures this finer phenomenon gives insights into creating strategies to mitigate hate speech. [**In Submission**]

Tracking Online Gender Based Abuse May'23 - Present

Advisor: Tarunima Prabhakar

- Tracking **online harassment** that Women Politicians are facing in the 2024 Lok Sabha elections. Performing a cross-platform study across YouTube, Instagram and Facebook to understand the engagement that women politicians receive on their social media. Mapping how does harassment spread across online platforms? What kind of content posted by women candidates is commonly targeted?
- Created a YouTube Dataset to track gender-based violence using a crowd-sourced slur list by Gender Right Researchers and Activists. String matching algorithms and NLP models were used to analyse text in Indic languages. [GitHub]
- Improved NLP systems to understand coded language (dog-whistle or double-meaning words).

Content Moderation for Indian Languages Jul'23 - Present

Advisor: Aditya Joshi

- Improving NLP systems for content moderation in Indian languages by breaking down the task into key sub-tasks, like hate speech, toxicity, sarcasm, sexism, cyberbullying, and gender-based violence detection.
- Adding context to models to reduce social biases helping in improving content moderation. Created a dataset of 3.2M Reddit comments and 133K posts, that reflected an Indian context, capturing socio-cultural aspects and linguistic diversity.

Talks

"Modelling Hate Speech on a Social Network"

- Social Network Analysis Class (CSE533) *Invited Speaker* [Link] Apr'24 (Ahmedabad University, India)

"Building NLP classifiers to detect Hate Speech"

- ACM Winter School on Network Science *Invited Speaker* Dec'23 (Ahmedabad University, India)

"Tools to respond to Online Gender Based Violence"

- Digital Citizen Summit *Invited Speaker* Nov'23 (Remote)

Achievements

Bosch Future Mobility Challenge | Semi-Finalist [Globe] Programmed and Engineered an **autonomous driving car** to navigate a miniature city. One of the 24 teams out of 118, and **only team** to represent **India** for the finals in Cluj, Romania. [Newspaper] [Video]

Outstanding Paper Award Workshop on Online Abuse and Harms at NAACL'24

Ingenious Hackathon 3.0 | Winner Open Source Track

Academic Service

Shared Task Organiser ICON'23 (Gendered Abuse Detection in Indic Languages) [Link]

Reviewer ACL Rolling Review'24

Teaching and Leadership Roles

Discrete Mathematics (MAT 101), Ahmedabad University *Teaching Assistant* Nov'22 - Mar'23

- Responsibilities included evaluating assignments, and helping students with the coursework.

Center for Learning and Empowerment (NGO) *Volunteer Teacher* [Globe] Jul'23 - Apr'23

- Taught 20+ secondary school kids mathematics for an entire academic year, at a tribal village in Jharkhand, India.

Reading Group, Ahmedabad University *Organiser* [] Dec'21 - May'22

› Organized a weekly Reading Group focused on reading recent and classical papers in Computer Science.

Programming Club, Ahmedabad University *Content Head and Event Organiser* [] Jul'20 - May'22

› Co-organized and curated events around Web Development, Machine Learning and Competitive Programming.

Google Android Study Jams Facilitator *Instructor* [] [] Dec'21 - Jan'22

› Conducted 3 sessions teaching Android Development in Kotlin to 40+ students.

Usable Pasts, Sustainable Futures *Student Volunteer* [] [] [] Jan'23

› Worked under Dr. Tejaswini Niranjana to understand how historically effective practices around Food, Clothing and Housing (*roti-kapda-makan*) can help mitigate climate change and public health issues.

Skills

Languages	Python, JavaScript, Elixir C/C++, Kotlin, GraphQL
Frameworks/ Libraries	PyTorch, Tensorflow, Nodejs, OpenCV, networkx, CameraX, Elasticsearch, Sequelize
Databases and Developer Tools	MySQL, PostgreSQL, Docker, AWS (S3, EKS, EC2), aws-cli, Kubernetes (kubectl), CI/CD, Android Studio, Firebase