5

…the tool is under development…

Cheshta Arora

5 **...THE TOOL IS UNDER DEVELOPMENT...**

Cheshta Arora

Zero

Imagine a user-facing content moderation tool in the making that uses machine learning (henceforth, ML) approaches to detect offensive and problematic content in three Indian languages: Indian English, Hindi, and Tamil.

Underpinning this story is a moment of failure narrated from my perspective as one of the team members involved in developing this tool. I probe, using broad brush strokes, the events of the last few decades in the history of postcolonial India to find a proper place for this moment of failure that was otherwise a trivial incident for the rest of the team.

The failure, however, introduces the ways in which new actors—like ML models—and new ways of knowing and doing enter our fields of politics.

The narrative consists of seven vignettes. If the reader is left probing, wondering, searching for that one declarative sentence that will say it all, this story has served its purpose.

One

There were long episodes of awkward silence forty minutes into the workshop. The activities we had planned for crowdsourcing a list of slurs, hashtags, and problematic content did not resonate with the workshop participants, a diverse cohort of activists, members of community-based organizations, academics, and individuals with strong digital presence.

This reticence on the part of the participants continued during the entire course of the workshop. A three-hour-long session had to be closed 30 minutes early. To sum it all up, the workshop was a failure.

I was part of an interdisciplinary team of computer scientists, social scientists, and activists developing a user-facing, browser-based web plug-in drawing upon ML approaches to detect hateful and violent content on social media.

Our presence on social media generates a vast amount of content, some of which is hateful and violent. A standard rhetoric of social media companies is to say that it is practically impossible to deploy human moderators to sift violent content from non-violent content (also, who will become human moderators to clean social media feeds, and for whom, is more than a technical question). Enter large-scale, industrial management of content: machine learning. In plain English, with machine learning, a lot of hateful and non-hateful content is fed into a computer and the computer learns, through pattern recognition, the difference between the two. There is, however, a twist.

Social media platforms have been consistently gobbling up different parts of the world to increase traffic on their networks. Despite profiting off the content uploaded on their networks, these platforms don't want to invest in moderation models that could detect content in these languages, especially the ones that are spoken in the majority world. In terms of providing services, they merely rent out the online real estate without taking any responsibility for what happens on their property.

With no intention to pursue nuanced alternatives, platforms are mostly satisfied with unimaginative approaches vis-à-vis problematic content on their network. They either censor it if it affects their imaginary community or feed on it since provocative, hateful content tends to bring them more traffic.

In response, the team thought of devising a moderation tool for non-English content while also envisioning new ways to think about content moderation. I, along with the entire team, was doing the dirty work for the platforms by devising tools to make their property safer. To be sure, the platforms of Web 2.0, the social media left over from the early 2000s, is rife with all sorts of hateful content—a constant dilemma of content moderation. But to start, we chose to focus on instances of hate speech, harassment, and violence perpetrated against persons of marginalized gender and sexuality who might also be situated at other intersections of caste, religion, or ethnicity.

Even a simple list of slurs was lacking in Hindi, Indian English, and Tamil. To crowdsource a list of slurs, understanding how different individuals and groups are attacked and how they respond to these everyday threats were the first steps toward building a tool.

With this workshop, we had hoped to trigger brainstorming sessions to narrow down the definition of harmful content, arrive at trade-offs vis-à-vis over-moderation, and map other, useful non-ML features into the tool. This, we had hoped, would become part of our co-designing methodology through which we could collectively design scaled-down machines that intervene into problems specific to our lives and of those around us.

The reasons for our failure were manifold: the workshop was conducted online on Zoom, and as first-time facilitators we were told we had failed in our attempt to reproduce an atmosphere of intimacy, understanding, and confidence—that the planned activities were too intense to sustain the attention of our virtual participants for three hours. All of these reasons and many other im/perceptible factors could have been at play. There were many obvious, perceptible mistakes that we could easily identify in hindsight. These mistakes contributed to the failure, but if there is any explanation for the failure at all, it lies elsewhere.

Two

During the start of our project in June 2021, our donor had published a blog post introducing it on their website. In that post, the tool was posited to reduce the problem of online violence faced by women and children everywhere and especially in India where it suggested that the problem of gender-based violence continues to intersect with India's centuries-old patriarchal society. In one stroke, this post erased years of postcolonial feminist work in the subcontinent which insists that the contemporary problems of the third world are as modern as the colonizing impulse of the first world. It also reflected an uncritical positing of tech solutions to problems that are more fundamental. It's as if we were still dealing with the postcolonial problematic of using technology to leapfrog into modernity. After reading the blog post, a feminist navigating the space of digital rights would comment that a true feminist would never undertake a project like this.

During the early months of brainstorming on the tool design, we witnessed the challenging phenomenon of online hate that had spread under the hashtag #sullideal. On Twitter, swarms of accounts began using the hashtag to harass Indian Muslim women; following this narrative of hate speech, we found at its heart an independently built application hosted on Github. The application was populated by publicly found images of assertive Indian Muslim women on social media. At the start of the app, the users were asked to click to "Find your Sullideal of the day."* Once clicked, the app would display a picture of a Muslim woman with the tagline "Your sulli deal of the day is" along with details about her social media handles. In a twisted communitarian spirit, users also had an option to share this on their own social media pages. The pop-up invoked the trope of "auctioning"—harking back to certain facets of Islamic history where women were purportedly enslaved during war. While the Github account that had uploaded the application was taken down the next

* "Sulli" is a derogatory term used by right-wing extremists to refer to Muslim women in India.

day, the targeted harassment and bullying continued for a good twenty days before losing its viral currency to other hashtags.

In response to this harassment, members of the women's movement in India wrote a letter condemning such actions. In their condemnation of this act, however, they uncritically lumped together several distinct topics: pornography, objectification, dehumanization, the sexualization of (Muslim) women; all were collapsed. So while the letter was weaved to cast a wide net of violence, to capture multiple issues, it also hollowed out the specificity of the problem. The women's movement had, apparently, already mapped the linguistic contours of this problem. Opponents of the women's movement pounced on this collapse of the pornographic and political, comparing #sullideal to fetish websites that presented images of Hindu women for Muslim men, further muddying the issue.

According to the dominant response by the women's movement, we were still caught in the problem of objectification, sexualization, and dehumanization where the difference between pornography and religious-gendered hate was placed on a spectrum. Pornography it seems was still a dirty word. Yet, we still don't know how to respond to young female Indian influencers on Instagram with Onlyfans accounts, digital expressions of queer sexuality on the web, Indian cam-workers on otherwise banned sites such as xHamster, amateur content of heterosexual swinger couples on MeWe, non-consensual sharing of images on Reddit channels, the proliferation of BDSM subcultures on Fetlife, or more generally, the horror that pornography is allowed on Twitter! The web is an ocean, and to think of a response for each one of these perversities would be a whirlpool.

The women's movement already knew the answers to the problem that it didn't understand.

This was the other side of the failure.

Three

Speaking of whirlpools, I am taken back to another day. I had recently joined the team and was just one month into the literature around content moderation. Its naive reference to problems such as misogyny, sexism, racism etc. had allowed it to recruit ML/AI as part of its rescue party. Before joining the project, however, I had also dabbled with feminist approaches to technology that opened up the problematic of feminism to machines,(1) dogs,(2) turtles,(3) monkeys,(4) mice,(5) and monsters,[6] to bodies not ending at our skin. In such articulations, the world was not neat and categorical but messy and relational. While machine-learning models can understand categories, a porous and messy world of monsters and turtles deeply entangled with the human and the machine was still an abstraction. Sadly, machines and our knowledge-making practices were not there yet. We march back to identity, misogyny, patriarchy.

Armed with these abstract, theoretical approaches to posthuman, decolonial, feminist machines, I was shamelessly demanding that we discuss the "narrative" that we want our tool to weave. Will this tool continue the abhorrent narrative of violence against women discourse? Would it be possible to code another narrative in this tool that can account for the complexities of postcolonial condition, locate a different causality to gendered, sexual violence that is other than an abstract, hollow notion of age-old patriarchy and its violent manifestations? This discussion, a team member heuristically suggested, would unleash a whirlpool. We had a timeline, we were accountable to our funders, and we could not afford to get lost.

My demand was shameless not only because it didn't take into account more concrete concerns of tool development, timeline, etc., but because of the inconspicuous high-handedness of this question. Where tech fundamentalists demand that their machines be most efficient, the theoretically informed social scientists, in their high-handedness, demand that the machine be most radical. Both techno-solutionists and the social scientists never forget to pay their regard at the pedestals of human supremacy and its cognitive capabilities.

As I got comfortable with the messiness of the process, this deference to human cognition would come back to haunt the workshop.

Four

During one of our one-on-one conversations with a queer feminist, on sharing the inhibition that a feature such as easy filtering of slurs could be appropriated by the troll army to block content from those at the margins, we were met with an indifferent shrug to suggest that nobody cares as long as the tool is useful for those who are targeted.

To assist us during our one-on-one conversations with activists and members of community-based organizations, we often use a MURAL board that gives a visual representation of our feature list. On the MURAL board,[7] a list is mapped onto a grid that indicates usefulness to the user on the Y-axis and ease of development on the X-axis. The contextual identification of slurs/problematic content which requires ML approaches is on the left—our hardest task—while easy filtering of slurs is on the far right.

During another conversation with a fellow-traveller-feminist engaged in building feminist tech in India, our MURAL board was interestingly flipped upside down:

— *"Yeah well, I wouldn't worry so much about moderation and specific detection of slurs.*

Giving an option to make complaints easier will be great.

I am 50-50 about the invoke networks for action feature.
Support networks keep changing.

Mental health prompts—bad, bad idea.

Detecting Virality is a good feature.

Archiving tools, perfect! One can have an entire history of incidents to make complaints. Especially if we can share it in our support network, and they can help archive every incident.

Tool can also suggest resources or options like: You could do this, this, this; report the post, whatever is shared on the platform, archive it, block the person, document, ignore, engage with the platform.
Simple strategies.

If I had limited resources. I would keep the filtering simpler and include archiving tools as one of the features…"

— *"Yeah well, I like how you have flipped our entire board!"*
one of us replied,
"but ML is a high priority issue for us because we are building it for the under-resourced languages (also this is something that we have promised to the funders)."

With each one-on-one conversation, it was becoming more and more clear that there was no desire for a more efficient, complex tool. Rather, the desire was for the simple features where one can work with the machine to mitigate the violence rather than strive to make the machine work in the background on behalf of the user.

While I was working on the project, I was also part of another that was collecting narratives of those with an active online presence who have been at the receiving end of online caste-based hate speech.(8) Our respondents pointed out pervasive discrimination that they face as content creators and how some of them who started during the early days of social media could not stand the monster that the web had become. During these conversations, I would come in at the end to ask participants to imagine a tool that could intervene into this experience. They didn't evoke an event of violence or extreme cases but instead they simply described everyday exhaustion vis-à-vis hate speech on social media. As one respondent said:

> — *"It's not that we fear it but we are tired."*

Trivia:
Do you know of all content on social media what percentage of it is hate speech? Statistically speaking, existing literature suggests that of all content on social media, hate speech is at an abysmal 2%.

Five

During the workshop, the participants were overwhelmed with the idea of crowdsourcing the list of slurs. They intuitively understood the inability of an ML tool to successfully distinguish between a problematic use of a slur, appropriated use of a slur, or the casual use of a slur. Then, the participants brought out the microaggressions that use humour, sarcasm, and stereotyping that form a crucial part of our politics on the web. They also pointed out that…

> *"…the definition of gender-based violence itself is limited; there is tech-based violence that uses IT to harass. Given the reality of the subcontinent, the circulation of images causes more harm than the text. Most people use Facebook and Instagram. Only an urban, English-speaking elite uses Twitter."*

"…yes, but would it be possible to crowdsource a list of slurs? How exhaustive would it be? There is a universe of language."
"…I don't want to undermine the importance of addressing everyday, individual fatigue, but the tool should also be able to intervene at a larger structural level. We should be able to use it to make some systemic change."

"…it's important to know the limitations of tech, we come across people with huge digital divides."

"…it has become a fad to develop tools."

The omniscient and omnipotent ghost of human supremacy had come back to haunt us. The new tool was expected to address all problems or nothing at all.

Six

As a team, we are invested in co-designing our tool with others who would be potential users. We are informed by those ML approaches to content moderation that insist on building diverse datasets while involving activists, community members, and individuals who are at the receiving end of violence as experts and annotators to arrive at a contextualized understanding of harm. We regularly insist that the data collected during the project be placed in open access repositories. The tool itself will be free to use and modify without any prior permission. This investment and ethical considerations are theorized within the existing literature on ML as ushering in better transparency, robustness, and accountability and to keep in check the unfettered ambitions of a scientist.

Transparency, robustness, and accountability, however, are innocent justifications. What's at stake is the fundamental question of relation. As sentient beings, what is our relationship with machines?

To come back to the question of failures, after our initial experience during the workshop, we decided to give up on the format of large focus group discussions in favour of smaller groups to arrive at some of the decisions that we must code into the model.

Seven

"Congratulations! You guys have built a perfect torture machine…"

This is how one of our annotators gleefully described the annotation task. We had invited 18 expert annotators (6 for each language) from the pool of activists, journalists, and community influencers that we were in touch with to annotate the data which will be used to build the ML model. Through their annotations, these 18 experts were teaching the machine the difference between a problematic and non-problematic post. By inviting 18 expert annotators, we wanted to capture the range of harms and how the same words work differently for different people.

However, the tyranny of the machine still demanded that different annotators agree with each other to an extent. To help with that, we had a set of instructions. We came up with three labels:

1. **Is it gendered abuse when not directed?**
2. **Is it gendered abuse when directed at persons of marginalized gender and sexuality?**
3. **Is it aggressive/explicit?**

The expert annotators had to look at a post through three different lenses: the text of the post, who it is directed toward, and the perceived tone of the post. They had to read a post without a context, imagine an average context/use, and imagine a best-case scenario. As per our calculation, they could annotate 40 posts in an hour.

They had to forget and imagine the best, the average, and the worst for each post every 1.5 minutes.

In the process, we had indeed created the perfect torture machine for our annotators.

Some of the expert annotators had to be told to not overthink and some were asked to be more expansive. That is, they weren't allowed to mark "good morning" as creepy, but they could mark "you are cute" as creepy under label 2 (i.e., when directed).

Despite all our efforts, we couldn't find that fine line between a creepy and a non-creepy post. Can a text in itself be creepy or is it the action of repetitive posting? Maybe the machine will recognize a pattern that tells more than what we can know. We are awaiting their agreement score. The tool is still under development.

Cheshta Arora is a collection of cells whose work traverses the ethnographic and the theoretical to find various expressions of future immanent in the present. To that end it has been interested in studying practices and utterances that remain incomprehensible to the present. In the domain of internet studies, it is exciting at the moment to chase the cypherpunk dream of privacy, decentralization, and distribution.

Acknowledgments
Thank you Jonathan Corpus Ong, David Scales, Miguel Rivera-Lanas, and Yasmin Curzi de Mendonça for listening to this story and providing thoughtful feedback and expertise.

The work was supported by the Centre for Internet and Society and Tattle through a grant from the Omidyar Network India. The tool was launched in July 2022 and can be downloaded from: https://uli.tattle.co.in/.

Endnotes

[1] Haraway, Donna J. “A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century.” In Simians, Cyborgs and Women: The Reinvention of Nature, 149–81. New York: Routledge, 1991.

[2] Haraway, Donna J. *The Companion Species Manifesto: Dogs, People, and Significant Otherness.* Chicago: University of Chicago Press, 2003.

[3] Haraway, Donna J. *Staying with the Trouble: Making Kin in the Chthulucene.* Durham: Duke University Press Books, 2016.

[4] Haraway, Donna J. *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York: Routledge, 1990.

[5] Haraway, Donna J., and Thyrza Goodeve. *Modest_Witness@Second_Millennium. FemaleMan_Meets_OncoMouse: Feminism and Technoscience.* New York: Routledge, 2018.

[6] Haraway, Donna J. *Promises of Monsters: A Regenerative Politics for Inappropriate/d Others.* New York: Routledge, 1992.

[7] The board can be accessed here: https://tattle.co.in/products/ogbv/

[8] The report from the project is available here: https://cis-india.org/internet-governance/blog/online-caste-hate-speech-pervasive-discrimination-and-humiliation-on-social-media