Reclaiming Our Voices: Imagining Community-led AI/ML Practices

Subhashish Panigrahi
O Foundation
psubhashish@gmail.com

A wide range of recent scholarship on machine learning and generative content-related technologies discusses how these technological advancements can preserve endangered languages. While this is a seemingly practical and timely question, we might want to take a step back and inquire who is responsible for endangering languages in the first place. Is technology to blame, or is it the age-old oppressive social practices? Furthermore, can technology alone address a systemic issue or the interconnected systemic issues at hand?

I would like to share with you the practices of communities engaged in building what we can refer to as "Slow Language Models" and "Small Large-Language Models."

We have witnessed an increased emphasis on the role of generative content, both textual and audio-visual, in recent months, despite machine learning and generative text having existed for several years. On one hand, the imminent danger of potentially losing half of the world's languages within a century is a real and pressing concern. On the other hand, hastily implementing AI solutions for many such languages could have catastrophic consequences. Among the at-risk languages, indigenous languages, in particular, possess rich oral traditions, and not all communities have adopted writing systems. Due to centuries of colonization and post-colonial social oppression by dominant neighboring communities, many indigenous communities are marginalized and divided along geopolitical and ethno-religious lines. Additionally, several communities with smaller populations use multiple writing systems either to assert their evolving identities or identities imposed upon them. For example, Konkani language speakers in India employ as many as five writing systems: Devanagari used by Hindu Konkani speakers in the states of Maharashtra and Goa, Roman script by Catholic speakers, Perso-Arabic script by Muslim speakers, Malayalam script by speakers in Kerala, and Kannada script by speakers in Karnataka. Each writing system is deeply linked to religious or geographical identity, if not both. Similarly, the written literature of the indigenous and endangered Sora language spoken in the Indian states of Andhra Pradesh and Odisha, is divided among three writing systems: Sora Sompeng used by those following the traditional faith system, Roman script by speakers following Christianity, and Telugu and Odia scripts by speakers in Andhra Pradesh and Odisha, respectively, which are the dominant and official writing systems in these states.

The design of deep learning models predominantly focuses on textual content. Most languages spoken in the world either do not have writing systems; some have multiple writing systems; and many are used less in writing, resulting in less standardization and less available textual content and data. Audio and video-based language models are heavily reliant on annotation and transcription and are, in other words, text-dependent. Tagging, annotation and transcription requirements for audio-visual content are labor and cost-intensive. The current large language models heavily rely on publicly available online data, which infers that the majority of the world’s languages are left out of the LLM race. This issue also predominantly reflects the limited documentation of the collective knowledge of communities, resulting in skewed and biased viewpoints. The developers of popular models today, such as ChatGPT, maintain opacity about the sources of their data while scraping large volumes of publicly hosted data without permission from the owners of the data. Rushing to scrape online data, particularly created by non-native speakers, exacerbates the issue of poor representation of many historically marginalized communities and complicates pre-existing community power dynamics. This problem adds ethical concerns and raises questions about data sovereignty and agency.

Prominent computer scientist Dr. Timnit Gebru, a pioneer in addressing intersectional issues in AI and machine learning, recently shared her experience of encountering very few Black people at AI conferences where thousands of attendees gather. Representation is not the sole factor, but it holds significant importance in establishing the foundation of any system, not just technology.

AI represents only the tip of the iceberg. When we consider any emerging technology, disparities are evident across the board. Race-based segregation and the resulting underrepresentation, as highlighted by Dr. Gebru, have led to limited access and participation in the technology sphere. Furthermore, factors such as gender, caste, ethnicity, disability, and other intersectional aspects play a pivotal role in determining who develops technology and who uses it, along with the extent to which users contribute to its development.

Returning to the original question: Can marginalized communities engaged in AI development rectify issues related to access, representation, and decision-making? It's a complex question without a single definitive answer. However, before attempting to answer it, we must delve into the current power structures that underlie machine learning and generative AI.

Large language models are often seen as social innovations, but this notion is largely a myth. Such models are primarily built using unattributed content, often "stolen" from various sources without proper acknowledgment. I use the terms "stolen" and "labor" carefully because this issue is personal to me. From 2011, I have dedicated countless hours to creating and editing Wikipedia articles and contributed over 25 hours of speech data, all released under Public Domain and other open licenses. I also invested significant time documenting this practice and training others. Many others, mostly with various privileges like myself and also passionate activists, have been contributing for years with a hope to expand access to factual information and an assumption of good faith that their donated content would be used in various
ways, including by researchers and developers creating language tools.

We must not overlook the fact that most community-based solutions arise from genuine needs and sheer frustration with existing systems that fail to meet those needs. In contrast, available large language models today are driven solely by profit motives, often presented as innovations, services to marginalized communities, or tools for access and participation. It is not inherently wrong to seek profit, but it becomes frustrating when profit-driven schemes are sold as solutions while overlooking their true motivations.

Large language models are overtly capitalist by design, prioritizing short-term, large-scale expansion without adequate accountability or ethical considerations. Technology, as we know it, has never been neutral; it has always been a tool, and more specifically, a capitalist tool. The shape and form of capitalist tech also changes dramatically in different contexts. What's concerning is that communities contribute to training tools like Google Translate with the hope of seeing their languages included in the next release. The same technology that exacerbates disparities also offers hope for improved access and representation. We are at a point where technologies like AI and ML are like polluted air, but stopping the breathing is not an option.

When existing financial power drives technology development, under-resourced language communities pay a heavy price. It is possible to generate a dense academic paper in English overnight using generative text models. However, I know of many languages where generating even a small amount of typed textual content has taken years.

We are aware that ChatGPT’s training data largely relies on Wikipedia articles in various languages in addition to other sources. The content of Wikipedia is not owned by the Wikimedia Foundation but by its editors, including individuals like myself. The Wikimedia community, which includes many indigenous and historically marginalized individuals, was unaware that their volunteer labor was being used, byte-by-byte, to train large language models for eventual commercial use in chatbots, a multi-billion dollar industry now. Ironically, Wikimedians must pay if they wish to use ChatGPT beyond a certain point.

We conducted a study on the web content monetization of two indigenous languages in India. The findings revealed that most content creators not only invest their volunteer labor but also their personal funds to create content. These communities have endured social oppression for years, and they believed that the digital world would provide them with a form of liberation. Why should it be acceptable for a corporation to appropriate their labor, without consent, to train large language models and then sell the output without attribution or compensation?

Another important consideration pertains to community priorities in relation to AI.

As Kenyan feminist E. Imungu Kalevera aptly said, "The prison of practicality is the place where new ideas go to die." Our society was not constructed solely on what some call objectivity; rather, it emerged from imaginations, dreams, and a combination of physical and interpreted experiences. Language has always been a tool for expressing these facets of human existence that are far from being objective. Science and technology do not always need to provide solutions; they can also serve as tools for documenting diverse experiences and imaginations.

Given that big data and fast AI are rapidly depleting our environmental resources, could “Slow AI” serve as a tool for advancing linguistic rights and abolition? The pace and priorities are ultimately determined by communities. For instance, the Kusunda community in Nepal, with whom I began interacting in 2017 while making the documentary Gyani Maiya, eponymous with the late community elder Gyani Maiya Sen-Kusunda, is reviving the Kusunda language by teaching local children how to speak it. They lack a writing system, but children in schools are taught how to write Nepali in Devanagari script. Due to limited resources, the only fluent native speaker, Kamala Khatri, and a local researcher, Uday Raj Aaley, are nurturing a young community of speakers. Their immediate focus does not involve creating a writing system or other complex aids. In the interim, Uday has compiled a dictionary, collaborated with Timothy Bodt to record hours of transcribed audio with the late Sen-Kusunda who was fluent in Kusunda, and I also had the opportunity to work with him. For many such languages, the first priority is to establish a working group within the broader community. If a language has not been actively spoken for decades, rekindling people’s enthusiasm will take time, and changing mindsets will require collective effort. Technology can certainly play a role in advancing community organizing.

It is high time that we flag tech solutions marketed as the saviors of marginalized communities. Abolition, community-led initiatives start where "savior tech" ends.

I acknowledge my privilege within my social context, from a caste and gender perspective. I must also acknowledge and apologize on behalf of my community for the social injustices inflicted upon neighboring Adivasi and indigenous communities. Reparation is a slow process, but it must begin somewhere. Like others, I too make mistakes along the way and remain open to criticism.

However, I do not intend to advocate for a neo-luddite perspective. Hope still resides within many underrepresented communities, and those with privilege can still contribute in a meaningful way. Surprisingly, these same communities have always held the answers, answers that we all too often ignored.

Where should we begin?

We observe two distinct paradigms in community practices related to emerging tech: (1) the radical approach, where communities build their own solutions, and (2) the hybrid approach, where communities volunteer to assist big tech in developing proprietary technology. Activists are vehemently demanding for fair and equitable treatment of communities and gig workers whose content, voices, bodies and knowledge(s) are used to create deep learning models. Nevertheless, individuals and groups still need to assert their presence on the web and various digital platforms, as language remains a political tool for assertion. Low- and medium-resourced language speakers are diligently and progressively building their own data to facilitate the training of future language models. This foundational work progresses slowly, with communities with written languages compiling word and sentence corpora, and speakers of spoken languages predominantly archiving oral history. Those outside of under-resourced language communities can act as catalysts by serving with their technical expertise and other resources to
help communities transition from a hybrid to a radical paradigm.

There are many issues speakers of underrepresented languages are working on solving, particularly in enhancing accessibility and utilizing AI as assistive technology. However, an eagerness to address the systemic issue of language loss using AI could have adverse consequences, such as contributing to child labor in cobalt mining in the Democratic Republic of Congo or displacing indigenous communities for data centers and e-waste landfills.

In 2019, researchers discovered that creating a generative AI model like Google’s BERT, with 110 million parameters, consumed energy equivalent to a round-trip transcontinental flight for one person. Consider the gross environmental impact if all solutions were to become AI-driven. On one hand, we must act urgently, as half of the world’s languages are at risk of fading away. On the other hand, the mere attempt to quickly solve the disappearance of languages using AI does not only erase the historical wrongs, but also any effort for reparation. Instead of covering up years of race, caste, gender and ability-based oppression by inventing AI-based solutions, privileged communities can contribute by consciously and ethically utilizing AI. It serves as a stark reminder that California, where it all began, was home to some of the largest number of indigenous communities across North America before colonization.

We can glean valuable insights from the media and tech practices of several low and medium-resource language communities.

Ho and Santali, spoken by approximately 9.6 million people in the east coast of India, and also in Bangladesh, Nepal, and Bhutan, are sister languages and have been my neighbors since childhood. I became involved with Santali digital language activists in 2014 and Ho-language activists in 2017, including publishers, writers, and others who aimed to create and access content in their respective languages. Thanks to affirmative action, literacy rates are improving for the native speakers who are descendents of those who struggled to end colonization among other nation-building initiatives, but were also oppressed by India’s dominant caste communities. However, only 10-30% of native speakers are truly literate in terms of their native languages and writing systems.

My community enjoyed the privilege of having a Wikipedia of its own, the Odia Wikipedia, in 2002, but it became dormant until 2011. A few of us took it upon ourselves to revamp it, but it is worth noting that we had the time and resources to volunteer for this effort. As the project gained traction and the community expanded, we believed we had figured it all out. However, we soon realized that we had a significant gender and caste disparity. Dominant caste cis-men were making decisions, leading to limited participation by women. To this day, there is not a single gender-queer person in a community that serves 45 million native speakers.

On the other hand, Santali has experienced steady and diverse growth. The Santali Wikipedia recently celebrated its fifth anniversary and has editors from four different countries. Naturally, there are biographical articles about numerous indigenous leaders in Santali. While existing socio-economic barriers have not completely eliminated the gender disparity, the community is actively addressing this issue. Noted Santali poet and lexicographer Maina Tudu who is also a Wikipedia editor, for example, taught Santali via her YouTube channel and a WhatsApp group during the COVID-19 pandemic.