Image: Midjourney prompt "community of humans, democratization, unity"

Guide to the Democratization of AI, ML, and NLP

Datasaur

Table of Contents

Artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) are revolutionizing the way society operates. We live in a world where there are discussions around sentient bots, ownership of AI-generated art, and liability for self-driving vehicles.

The technological developments are cutting-edge, and they're happening fast. As technology advances at a rapid rate, questions are also being raised around ethics.

One ethical topic that seems to be largely overshadowed is democratization. This guide will look specifically at how democratization and ethics are coming into play for AI, ML, and NLP.

What Does Democratization Mean for Technology?

Democratization refers to appealing to, representing, or making something available to everyone. When it comes to AI, ML, and NLP, this means making sure that technology and its advancements represent—and are available to—the masses.

How Does Democratization Come Into Play for Technology?

We're living through a technological revolution. And this is exciting, but it also means we're walking through uncharted territory all the time. We're having to answer questions and tackle ethical issues that we could never have imagined.

The issue of democratization is part of the ethical conversation that often gets overshadowed as more widely visible discussions play out around things like liability issues for autonomous vehicles. Yet democratization underpins a lot. How accessible, representative, and available technology is is pivotal for dictating the way that technology develops.

For example, if datasets behind the language models that power tools like DALL-E 2 and Midjourney overwhelmingly represent a select few languages and cultures, can those datasets—and the models they output—truly represent other cultures? Let's dive into this a bit more.

Democratization and AI-Generated Art

DALL-E 2 and Midjourney are AI art-generator tools that create unique images based on prompts given by the user in natural language. The tool then uses the prompts to generate semantically plausible images, with options for customizing things like the artistic styles and aspect ratios.

This is possible because the tools use NLP to train the deep learning models.

The widening access to these tools marks a step towards democratization, but questions are also being raised around democratization in different ways. For example, how are different cultures and languages represented in the datasets? How widely accessible are the prompts for users of different languages? How much weight should these questions carry for those developing the tools?

In practice, there are tangible ways that the representation issue plays out. For example, inputting keywords like "Indonesian" in AI art-generator tools will often return sketch-like textures and display a more primitive culture because Indonesian artwork isn't well represented online. As a result, the images that are produced are often outdated or reflect different periods in history.

Democratization and technology are inextricably linked, in ways that many continue to overlook.

Graphic: AI-generated art for "Indonesian"

The prompt in MidJourney "Javanese Doctor Who wearing a white and a black blangkon in front of the Tardis" returns a convincing image. But note that the texture is visibly that of a painting done with a brush. A prompt for "12th Doctor in front of a tardis" returns a more crisp image although it is still visibly something one would make in an art-related application.

A TL;DR of The History of AI, ML, NLP, and Democratization

Historically, advancements in NLP for English and Mandarin content have eclipsed all other languages. These are the languages that are represented the most strongly in the datasets, and they are also predominantly the mother tongues of the researchers and developers working on the state of the art.

Other languages, often called "low resource" languages, are largely underrepresented in the data. This means that we can't build the technology and datasets for these lower resource languages as easily because the information simply isn't available.

Now, as ethical questions are being raised around technology, questions about democratization naturally follow, too. Many large companies are opening access to their tools and datasets, and more questions around the topic of democratization are starting to crop up.

Graphic: AI generated art

[1] A note on this term: "Low resource" is largely a misnomer, but it is also a phrase that has become so widely accepted that it is the accepted phrase to d

Some Risks of Not Democratizing Technology

There are many potential risks that come with not democratizing technology. Perhaps the best way to demonstrate the risks and their impacts is to use some real life cases:

1. **Facebook's Voice Traffic**

   A few years ago Facebook found out that 50% of their total voice traffic came from Cambodia. This is not proportional and it was a surprising find. It happened because Cambodian users find it too time difficult to use the Khmer script on their devices (especially as smartphones became the norm), so users would turn to voice messages instead.

   Why does this matter? The Cambodian language is not well represented in AI and NLP developments, and if developers of these tools don't have exposure to the language, they don't see where the issue lies and the technology isn't accessible for the Cambodian people.

2. **On Stochastic Parrots**

   A researcher was fired because of a paper they were trying to publish about the development and potential harms that could result from large language models. The company they worked for was not criticized explicitly in the text, but it did discuss the pace, risks, potential environmental impacts, and financial considerations of language models.

   The fact that there was company pushback towards the researcher highlights the fact that the ethics and access questions around these technologies are highly sensitive.

3. **Mistranslation Leads to Arrest**

Someone in Palestine was arrested because what they wrote online was mistranslated by the social media company. This is the result of lower representation for languages like Palestinian Arabic in the language model datasets. Datasets in low resource languages often lack high quality data, which hinders the development of good machine learning technology for that language.

In today's world (where you can be liable for what you write online) a faulty translation can become dangerous.

Why do Lower Resource Languages Often Lack High Quality Datasets?

There are limited resources in our world, and much of the research and development so far has focused on the higher representation languages like English and Mandarin. This means that shortcuts are often made when capturing data for the lower resource languages. This may be because the researchers don't truly understand the language, or because developing a fully accurate dataset requires the right people and the right technological infrastructure, which can be a challenge.

As a result, when people develop tools for low resource languages they often use algorithmic systems to collect the data, but the resulting dataset and technology might not be accurate or representative of the language in practice.
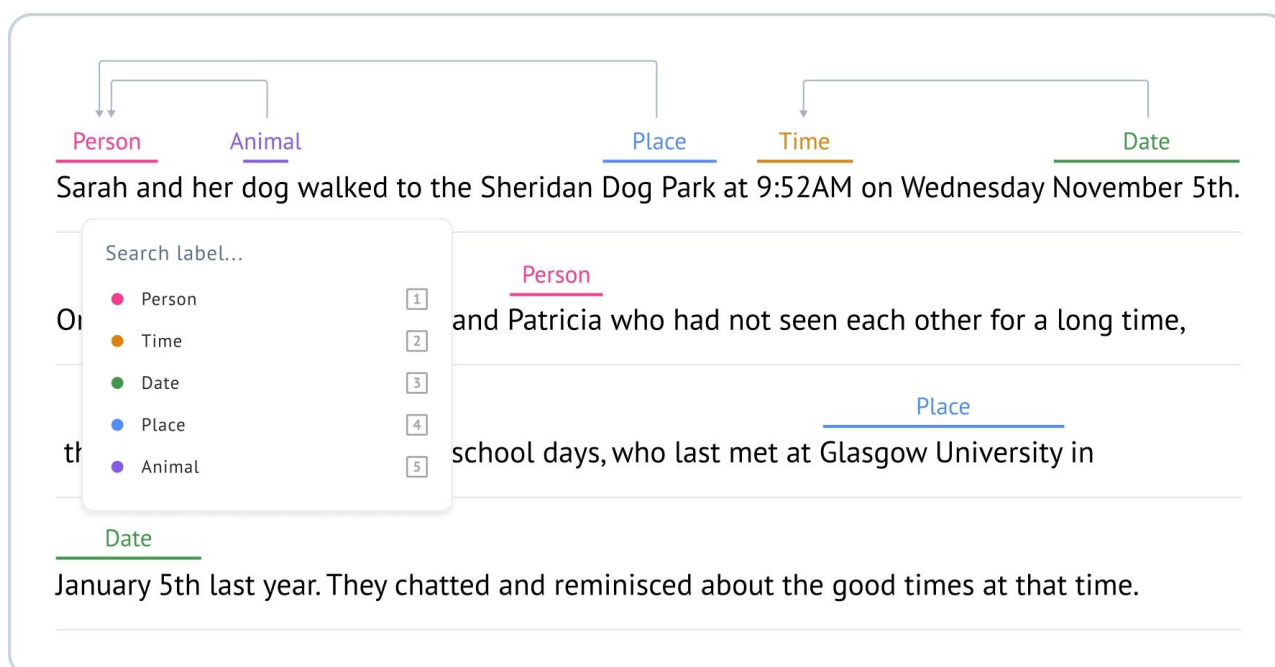
Graphic: Data labeling

Obstacles in the way of Democratizing Technology

There are several obstacles that can get in the way of making technology and data truly representative and accessible for all. These include, but are not limited to:

1. **Infrastructure**
   One of the most important factors is not having a versatile tool to help people develop democratized technology. If you have a sophisticated tool to help, it's more efficient to collect the hard-to-get datasets.
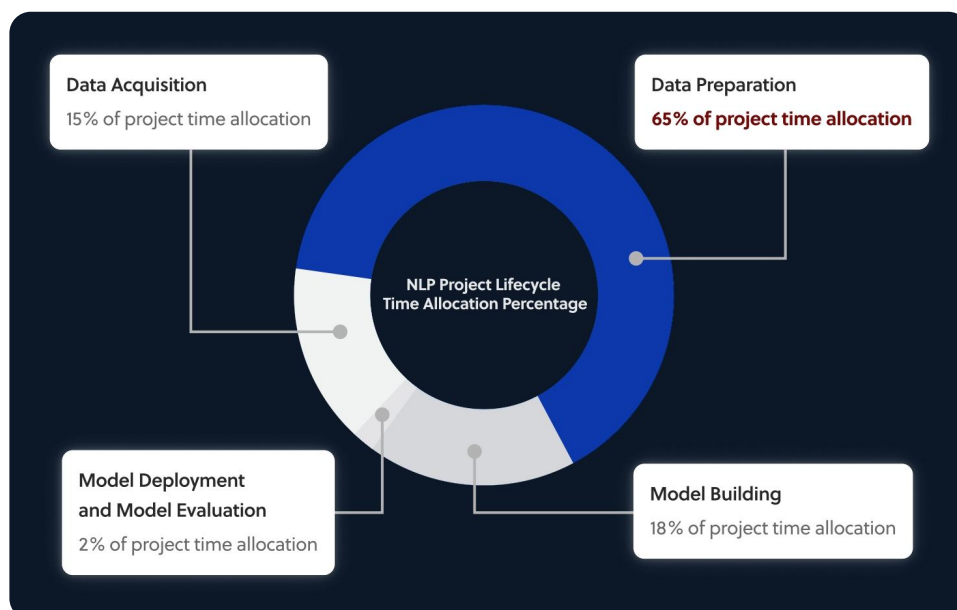
2. **Datasets**
   To build truly democratized technology, the data behind it needs to be accurate. At the end of the day, a model is only as good as the data it is fed. Remember the analogy "garbage in, garbage out." This is a major difficulty, as it can be expensive, challenging, and time consuming to get that data. The three main pillars here are:
   a. Collecting the data - Getting data that is accurate and representative.
   b. Processing the data - Data preparation takes a substantial amount of time, and poorly cleaned data can destroy the resulting model.
   c. Training the model itself - Having the monetary and human resources to develop the model, make changes, and improve upon it can be a massive barrier.

3. **Business incentive**
   Technology currently represents—and is most accessible to—the privileged few. This is in no small part because the major companies that are developing a lot of the cutting-edge technology are naturally driven by their bottom line. This is a complex and often misunderstood piece of the puzzle that could be an entire eight-page piece on its own.

Data Acquisition
15% of project time allocation

Data Preparation
65% of project time allocation

NLP Project Lifecycle
Time Allocation Percentage

Model Deployment and Model Evaluation
2% of project time allocation

Model Building
18% of project time allocation

Graphic: NLP Project Lifecycle Image

Potential Pitfalls of Democratizing Technology

Democratization can be a double-edged sword. Again, there are many ways that this can play out, and there is some debate around whether democratizing technology is necessarily the best path forward.

One of the potential pitfalls of widening access to technology is that if more people have access to the technology, there are more people who can deliver harm with it (intentionally or otherwise).

For example, YouTuber Yannic Kilcher trained an AI language model using toxic content from 4chan's Politically Incorrect board, which is notorious for racism and other hateful speech. Kilcher created ten bots and set the AI bots wild on the board. It took nearly two days for people to discover the bots, with some bots taking longer than others to be picked up.

The bots were able to interact with users in 4Chan, and in just 24 hours they wrote 15,000 posts that included a vast amount of racist and hateful content. Nicknamed "GPT-4chan" (after OpenAI's GPT-3), the model picked up hateful words and an overall tone that Kilcher said encompassed "offensiveness, nihilism, trolling and deep distrust."

Training a model with hate speech can be beneficial in that we can study people's tendencies, which can in turn help us to develop safeguards against that type of hateful behavior online. However, releasing the model with open access and letting it interact with people can inevitably deliver harm.

As access to technology widens, the safeguards around that technology must be bolstered up, too. Otherwise—as in the case of AI models like GPT-4Chan—those who are tech savvy can download the model and plug it into Twitter, for example, and use it to spew hate speech at scale.

Looking to the Future

Technology is often coming before ethics. Technology is advancing fast, and we are only now starting to understand the harms, risks, and why ethics are important. As we develop technology and learn more about its impact, we must also develop safeguards around ethics in AI, NLP, and ML.

There's hope and optimism to be had, though. There are a growing number of people who are advocating for more ethical safeguards around technology. And while the privileged few are still those that are choosing where technology grows and where energy goes, there are a lot of movements towards open access that take a big step towards democratization.

Massive Language Models and the Future

A lot of energy and development right now is focused on massive language models. Large language models are trained on vast tokens and we often don't know the details of the datasets for them.

This comes with inherent risk because we don't know how clean or toxic the datasets are. We also don't know which languages are represented, so we have no true way of knowing why tools like GTP-3 respond to prompts in the way they do.
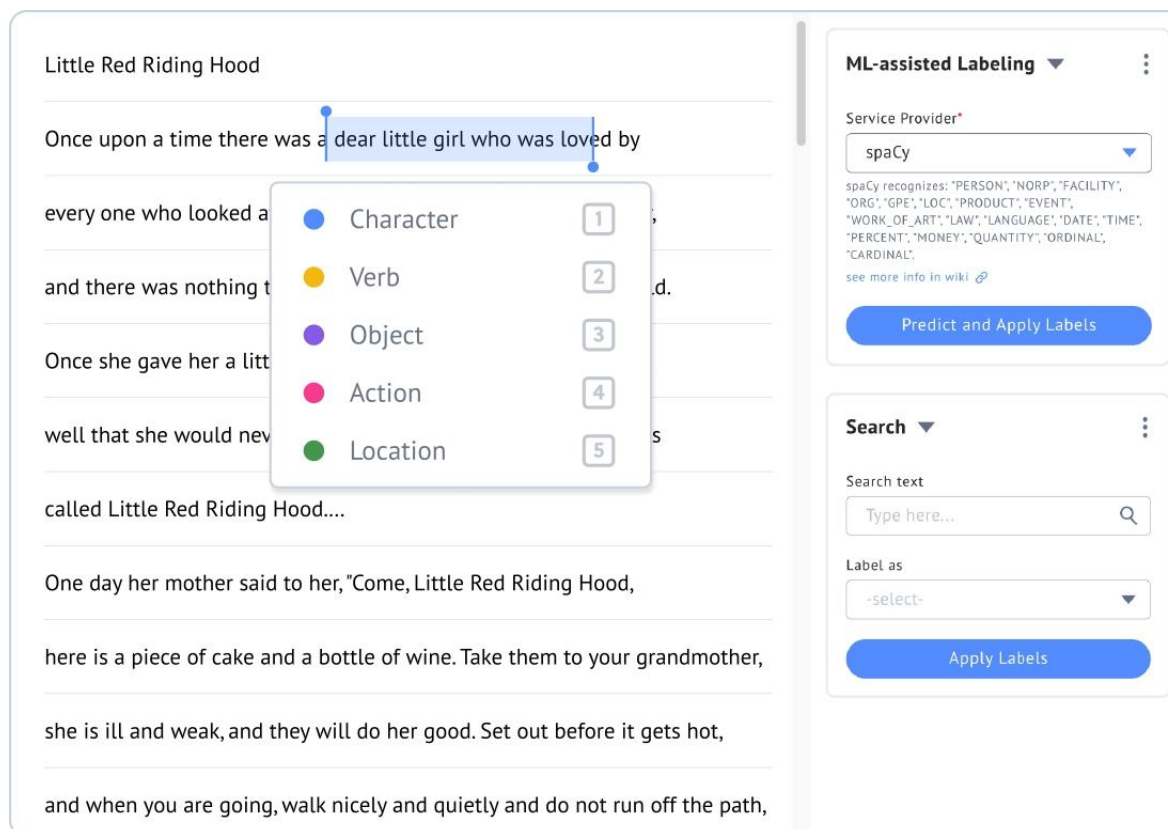
Graphic: NLP labeling

On the flipside, there are a growing number of massive language models like Meta's OPT and BLOOM that are opening access and widening representation. Meta's massive language machine learning model can translate between 200 languages and is trained on data obtained with "novel and effective data mining techniques tailored for low-resource languages." Similarly, HuggingFace's BLOOM has been trained on 46 different languages including code, and is built with open access at the core. These initiatives symbolize a big step towards wider access and wider democratization.

We still have a way to go. We need companies to continue to be transparent with the models they create and the datasets that train them. Meta gained a huge amount of publicity for opening access to their massive language model, which means that they have a vast number of people inputting data and conducting research. Many companies will not have as much traction or access to people, though, so this may not be scalable in a wider sense.

This will be an ever-evolving conversation. As new technologies spring up, so will new questions. It's very complex and the questions are large. But it's important to keep having these conversations. The more people are educated on what is happening in technology and what they have access to, the closer we are to fair democratization of AI, ML, and NLP.

Datasaur's Stance

At Datasaur, we believe we should democratize technologies to make sure that every language in every geography can benefit.

We want to work to make sure that everyone around the world can benefit from the NLP revolution. One way that we accomplish this is by supporting all languages on our platform. We realize that a lot of languages don't have basic NLP models for things that we take for granted.

For example, in the US we know that Wells Fargo is a bank and organization, but we don't know what the equivalent terms are in Venezuela. If NLP tools aren't built to support different languages and cultures, the resulting models cannot understand those sorts of things. The tools we built reflect our stance. We build tools that are robust and allow people to build labeled datasets quickly and efficiently, which in turn benefits people and allows for the development of—and research around—technology for all groups.

For lower resource languages that don't have as many users, a sophisticated tool will help users to build a technology that can uplift those users.

Conclusion

Technology is fundamentally changing the way our society operates and the way our companies do business. NLP is right at the cornerstone of that revolution. The more people are able to use natural language to interact with computers, the more things change. This means that we're living through a phase of rapid and significant technological transformation.

It's fascinating to watch the way that technology continues to evolve. And it's fascinating to see that questions around ethics are starting to take center stage. People are only now starting to truly understand why ethics and democratization are important in the worlds of AI, ML, and NLP. This will continue to be a conversation as long as technology continues to evolve. And that's a good thing. The more that we talk about it, the more educated we can all become.
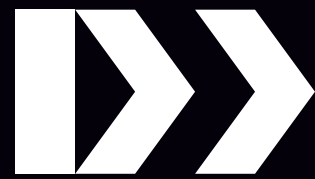
If you'd like to learn more about Datasaur's stance, our goals, or how we can support your NLP needs, please reach out and we'd be happy to talk.

Find out how Datasaur can help your business

https://datasaur.ai

Schedule a demo