Datasaur

Guide to Data Labeling Best Practices

August 2022 | Guides

01 Introduction

So you're looking to deploy a new NLP model. Perhaps one already exists and your goal is to improve precision or recall. You've tried multiple models, tweaked the parameters, and it's time to feed in a fresh batch of labeled data. Your company has real-world data readily available, but it needs to be labeled so your model can learn how to properly identify, classify, and understand future inputs. The quality of your labeled data directly affects the performance of your machine learning application, so it's vital to have a solid process for outputting high quality, accurately labeled data.

This guide will examine options for labeling that data and offer insight into how Datasaur can help you label data efficiently. We'll cover the following best practices:

1. Sourcing data labelers
2. Establishing a multipass labeling process
3. Setting up comprehensive guidelines
4. Scaling your data labeling
5. Using the best tools
6. Iteration and continuous improvement

02 Why is Labeled Data Important?

Before we dive into best practices, it's important to talk about why it matters to have the best practices in the first place. So, why is it vital to have robust processes in place for labeling data?

ML is a "garbage in, garbage out" technology. The effectiveness of the resulting model is directly tied to the input data; data labeling is, therefore, a critical step in training ML algorithms. To output quality models, you need to input quality data. Since machine learning and NLP are becoming pivotal for more and more industries, the task of data labeling is also becoming increasingly pivotal.

And it's a big task. Data preparation (think organizing, cleaning, and labeling data) can take up 65% of an NLP project's time. This time can feel like a massive sink when you're racing to properly structure the data to train and deploy powerful models. Knowing this, it pays to optimize your data labeling process and be able to scale efficiently. In turn, this lets your team output high performance models faster.

Labeling Options: Crowdsourcing or In-House

There are a few options for labeling your data, and deciding on the nuts and bolts of how to actually label the datasets is one of the most important decisions you'll make. You need to find the right approach for labeling the data before you can establish the best practices for that approach. So let's take a look at the options available to you:

Crowdsourced Labeling Vendors

Crowdsourced vendors (like Mechanical Turk) and labeling services (like iMerit and CloudFactory) will take in your data and send it back labeled. The benefit to this is that you have access to a large team of labelers, though it's important to know that those labelers are paid on a per-label basis, and quantity can be prioritized over quality. Also, since these labelers aren't in-house, they won't necessarily be familiar with the intricacies of your industry, company, and data, which can—again—compromise quality for your datasets.

In-House Labeling

You can also build labeling tools and processes in-house. In-house labeling could take the form of programs like Microsoft Excel, building your own tools, or using an open source option. Using your own tools can feel like an appealing option for the sake of centralizing costs and workloads, but it can also be inefficient, lacking in configurability and advanced labeling capabilities, and in the long run it can end up sinking far more costs and team resources. Plus, by having a labeling team in house, you have greater control over the label quality itself, particularly as it comes to very complex text that is specific to an industry.

Labeling Tools

You can also use NLP labeling tools (like Datasaur) to help label your data. This gives you access to more customizable and configurable tools built specifically for NLP labeling. Datasaur will also take the time to understand your unique data labeling needs and processes, and can help you to establish best practices for your workflow. We also partner with iMerit, CloudFactory, Isahit, and Prosa, if you would like to have labeling services (e.g. they do the labeling for you).

Choosing between in-house labeling tools or tools like Datasaur is a big decision, and can signify taking the next step in your NLP adoption.

Data Labeling Best Practices

Source the Right Data Labelers

The data labeling process needs human labelers to annotate the raw data with the corresponding labels. To annotate the data successfully, you need to source data labelers, and specifically the right type of labelers for your needs. Typically, labeled data for machine learning requires one of the following data labeling team models:

1. **Crowdsourced** - Small tasks are delegated to a large number of people for labeling. This can be a good option for basic, objective, non-specific labeling tasks (like names of cities).
2. **Non-specialized outsourced** - These are vendors that can deliver data labeling services with more accuracy than crowdsourced labelers can. However, they are not specialized by industry or topic.
3. **Specialized outsourced** - These workforce are highly trained in specific areas. They're able to provide much higher quality data labeling services, as well as often being able to help estimate labeling project workloads, timeframes, and more.
4. **In-house** - These labelers are on your payroll, either full-time or part-time. Their job description may or may not include data labeling. This gives you the highest level of control over the quality and accuracy of your data labeling, since you can maintain consistent communication, review, and training.

Choosing the best team model is crucial for streamlining your data labeling efforts.

Establish a Multipass Labeling Process

Establishing a multipass labeling process lets you get to a labeling ground truth. Every human being is fallible, and even the best labeler in the world can make mistakes or introduce subjective bias in data labeling. To mitigate this, a common practice is to have two or more labelers labeling the same data. For some projects, a majority consensus is sufficient for determining a labeling ground truth. Others will require unanimity and a discussion around each disagreement.

This goes hand in hand with establishing a robust QA and review process, which is critical for making sure that things run smoothly and labeler error is mitigated. You'll want to make sure you have a process for things like inter-annotator disagreements, as well as for reviews at the team, individual, and project level. Using a labeling tool with robust workforce management features will help streamline this.

Set up Comprehensive Guidelines

As the old saying goes: "by failing to prepare, you are preparing to fail." A common stumbling block in NLP is a lack of specificity and preparation when setting up projects. For example, setting an NLP goal to remove "inappropriate content" online will be difficult. It's ill-defined and has far too much gray area. Labelers will have questions around how to handle jokes, idioms, politics, sarcasm, and so much more.

To avoid cumbersome review processes and labeler confusion, you need to have well-defined guidelines. This will include setting up processes and definitions for edge cases, so that the product, engineering, and labeling teams can all make sure the right data is being fed into the model.

Scale Your Data Labeling Processes

Your data labeling team will never be static. As the complexities of your data labeling needs deepen—and as you get higher volumes of data—you need to scale your team to match. Data labeling for machine learning is iterative and time consuming, and you need a team that can handle your project complexities.

Scaling your data labeling process usually means including more people. There's more to it than that if you want to follow best practices, though. As you add more people, more things will need to happen to keep your processes robust. For example, more people means more opportunity for error, so you will need solid processes for measuring and managing efficiency, productivity, and quality across the board. You will likely need to add more reviewers, more people will require more training, and you will want to analyze the workforce to ascertain who's best at which types of tasks so that people are assigned as appropriate. This all goes hand in hand with establishing effective communication, checking labeling standards across the team, and choosing processes or tooling that match your needs.

Again, a robust labeling tool with workforce management features will help you dig into team, project, and individual level reports. This lets you see who is performing best in which areas, where the discrepancies are cropping up (and between whom), and where there are opportunities for more training to continue to improve team—and project—efficiency.

Use the Best Tools For Your Needs

There are so many tools and data labeling companies you can choose from, and finding the right tools is perhaps the biggest consideration for your data labeling. Choose AI labeling tools that streamline efficiency, work well with your team, and work for your specific use cases. Some things to consider when you're choosing the right tools:

- Security and data protection - Look at certifications such as SOC 2 and HIPAA. Also consider data storage policies and things like user access controls. When you're working with sensitive, personal, and/or non-anonymized data, military-grade security is important. Look for features like PII anonymization and end-to-end encryption.
- **Deployment options** - Consider how the tool will be deployed and where it will run from. The most common deployment options include: VPC and on-premise, hosted options (like Datasaur hosted on AWS), and deployment on a public cloud. Again, security can come into play here when considering the best options for your company.
- **Labeling features** - Consider which labeling features are essential for your processes. Do you need hierarchical labeling? Do you need ways to label individual entities and long form text without the UI becoming cluttered? Do you need ways to draw relationships between entities? Do you need an audio labeling tool that can handle speaker diarization, noise mitigation, and transcription? Whatever your needs are, make sure your tools can meet them. (And trust us, there are tools out there that can meet your needs, no matter how complex.)
- **Configurability/customization options** - If your labeling needs are complex, you need tools that are configurable. Many tools—especially open source options—aren't very configurable without tremendous effort, but options like Datasaur are. With Datasaur, we can customize the interface and configure your data labeling workflows to suit your needs.
- **Workforce management features** - Being able to effectively manage your projects, teams, and labelers at every level is vital. Make sure you're able to run QA, check progress, flag (and resolve) inter-annotator disagreements, and pull up reports easily and efficiently to analyze your progress.
- **Integrations** - Integrations can make or break your workflow, causing parts of your project to be much more tedious than necessary or by automating the tedious aspects away. See if you can plug in your existing model via API or use open-source label libraries like HuggingFace or spaCy to label the bulk of your data automatically. Check for automatic project creation and exportation. These integrations allow your team to focus on the actual task at hand, rather than monotonous, pointless tasks.

Iterate: Think Continuous Improvement

Training a model for machine learning is iterative. Your overall data labeling processes should evolve with you, and so should the data labeling for each model you deploy. As each ML model is deployed, you can continue refining its accuracy by refining your labels and implementing a human-in-the-loop model for continuous improvement.

1. **Improve your data labeling processes** - For example, if you need to label 500,000 documents, start with a small subset. Review that batch of data and make sure the labeling meets your standards. It's highly unlikely that the first labeled batch will be perfect, and adopting an iterative approach will save time, money, and resources as you move forward.
2. **Iterate on each ML model's labeling** - Once the model is deployed, we can continue to refine it. Human operators and labelers can step in when the model fails or could be improved. In edge cases, humans can perform tasks and log the case and train the AI model on that edge case specifically. Continue to pick up the cases, log, and improve the data as the model is deployed and improved.

03 Datasaur to the Rescue

At Datasaur, our mission is to build the best data labeling tools so you don't have to. Our text and audio labeling tools are designed with the data labeler in mind. We understand your labelers deserve an interface attuned to their needs, providing all necessary supplementary information at a glance while keyboard shortcuts and hotkeys keep them working as efficiently as only a power user can.

Datasaur is the most robust, customizable NLP labeling tool on the market. We are also dedicated to continually building additional features learned from years of experience in managing labeling workforces—and, honestly, learned from you (when you put in a request, we listen, because you understand your needs best!). Reach out to us to set up a custom demo—this will let you see how Datasaur could work for your labeling workflows and data specifically.

About Datasaur

Datasaur is a private LLM provider and data labeling platform designed for companies to build their AI ecosystem with ease and efficiency. It assists organizations and universities in setting up custom LLMs and annotating data more efficiently and accurately through automation, quality control, and human-in-the-loop workflows. For more information, visit www.datasaur.ai.

Schedule a demo