How to Ensure Quality Data for AI Algorithms

We all know false positives in machine learning can be costly. And while we also know that high quality data is imperative to the success of your algorithm, in some cases, data quality is even more critical than others. For example, a false positive in an autonomous vehicle or biomedical algorithm could mean life or death, however, in the case of an e-commerce chatbot, it may just result in poor customer service.

Since the weight and severity of a false positive differs across verticals, it's important to define the level of data quality and domain expertise needed to train your algorithm, as a part of your training data strategy.

Should I use an existing dataset, or do I need to produce my own?

Data is the most crucial element of machine learning. Without it, training an algorithm would be impossible. Depending on your use case there are a number of **open-source datasets** available. However, if the subject matter of your algorithm requires specialized data, you may need to produce your own dataset.

Does my vendor have an established quality assurance process?

Before choosing a training data provider define a list of quality expectations. Sama has worked with 25% of the Fortune 50 to ensure quality SLAs of 95% and above. An established quality assurance process may include but is not limited to:

- **A Quality Service Level Agreement:** This will depend on your algorithm, but it's crucial to communicate this to your vendor and understand how quality requirements will impact costs. For example, achieving a 99% quality SLA will require more effort than obtaining 95% quality.
- **Precision Levels:** This is vital specifically for vector segmentation projects and plays a major role in efficiency. Precision levels have a positive correlation with the time spent on a given task. Determining this in the beginning of a project will help you establish an accurate timeline for your project.
- **QA Bots (automation):** This is a method of automated checking for quality requirements such as label inclusion/exclusion (i.e. label 1 and 2 can not coexist) or label threshold rules (i.e Nothing < 10 px should be annotated). This impacts the speed of the quality analysis process.
- **Feedback Loop:** Receiving a real-time report of tasks, including errors, will result in greater stability of data in an accelerated time frame.

How do I validate the quality needed to train my algorithm at scale?

In the last decade, Sama has submitted over half a billion tasks with tens of billions of labels while training data for machine learning algorithms. During this time, we've learned that pilot projects are a cost-effective and highly efficient way to validate what's needed to train AI algorithms at scale.

Whether data annotation is done internally or outsourced, pilots help you determine what's needed to accelerate your ML pipeline and ultimately quickly develop world-leading AI. Among others, a pilot can provide evidence-based answers to the following questions:

- How long will it take to train and validate my algorithm?
- What is needed to ensure quality thresholds?
- What is the estimated cost to obtain high-quality data at scale?