Achieving AI ROI Through Data Quality and Diversity

TWO ENTERPRISE USE-CASES

Co-created by

Emerj Artificial Intelligence and **ClickWorker**

emerj
The AI Research and Advisory Company

clickworker

Introduction

Companies invest in something if they think it generates value, and AI development is no different in this regard. Myriad sources tell us that AI adoption is becoming more ubiquitous, which would seem to indicate that the people running these companies understand the business value and ROI.

Per research conducted by McKinsey, business leaders report allocating 20 percent or more of organizational earnings to artificial intelligence. Moreover, McKinsey states that said leaders project AI investments will continue to trend upward.

Christian Rozsenich, CEO of Clickworker, talks about his company's experience:

"AI technology has become more important for a lot of clients of ours, and many of these clients are investing into developing new AI technologies and processes."

Christian Rozsenich
CEO of Clickworker

clickworker

Mr. Rozsenich cites the increase in AI in consumer products, enterprise software, and automation as evidence of AI proliferation.

Perhaps the most critical element of good AI modeling - and thus proper deployment - is well-sourced training data. Sans such data, a good AI model - and thus successful deployment - is likely impossible. Clients are also demanding good and diverse data, says Rozsenich.

> ***There is a large need for data which need to be accurate [and] diverse in order to make development successful.***

The quality of an organization's data directly influences the quality of the AI model, and in many cases, the organization's decision-making capabilities. Enhanced decision-making capabilities inevitably add value to an organization. In a research paper, Gartner states this fact directly: ***"Improved data quality is a primary source of value for many IT-enabled business initiatives."***

Such demand drives the business model of Clickworker, a company that aims to provide quality training data for its clients.

While obtaining a vast amount of source data may not be too difficult, getting data that is of good quality and diversity is. The reasons for such are various, but Mr. Rozsenich cites issues created by complex legal data privacy frameworks like GDPR in the EU, which has rendered much of the previously-acceptable data unusable for training new AI models. It is also difficult for data scientists in certain industries to source good data. Rozsenich also states that certain types of organizations may have difficulties acquiring data for other reasons. For example, startups and small companies must often acquire quality training data on a limited budget or don't yet possess this data.

When asked why a company may want to look externally for help with this process, Mr. Rozsenich says:

> ***"In many industries, it is very difficult to acquire the necessary - and often specialized training data … So companies need someone who can help them source the data, ensure quality data, and ensure that the data is legit."***
>
> **Christian Rozsenich,** CEO of Clickworker

In this paper, we will examine two critical elements of good data - quality and diversity - via two of Clickworker's typical use cases: facial recognition and voice recognition. We will also expound upon what constitutes good data quality, along with the challenges of acquiring said data. We'll also look at what steps one leading data enrichment firm, Clickworker, did to overcome these challenges and add value for its client.

USE CASE #1

The Value of Data Quality and Diversity In Facial Recognition and Authentication

Business Challenge

A large manufacturer of silicon chips approached Clickworker for help with acquiring training data for their facial recognition software. The manufacturer wanted to embed facial recognition technology into their devices for user authentication.

Rozsenich states that gathering good source data and training such a model is challenging. The main reason is that facial features aren't static or uniform. A person's face changes over time due to biological factors such as aging, for example. Our facial expressions often change voluntarily and involuntarily, so the training data must take this into account. Finally, we also have different skin colors and pigments, which can throw an algorithm off, if not accounted for. All of these dynamics must be taken into account via a robust dataset to comprehensively train a ML Algorithm.

Adding to the challenge, Rozsenich says, is that the quality of the user-provided data differed because of factors such as device resolution and recording environment. Someone snapping a photo with the latest iPhone in front of a green screen will look vastly different than someone using a 5 MP camera in bad lighting, for example.

To overcome the above mentioned difficulties, Mr. Rozenich states that the company had to acquire, refine, and enrich a diverse dataset. A ***"diverse dataset"*** simply means a dataset with the necessary variety to train the model on a full range of potential scenarios, including ***"edge cases."***

Turning Data into Value

> ***There is a large need for data which need to be accurate [and] diverse in order to make development successful.***

To ensure robustness in its facial recognition data, Clickworker explains that thousands of its users photographed their faces ten times at different angles and with different expressions.

The company website includes an image depicting the angles: **four in the front** (neutral, middle, above, below), and **three to the left and right** (above, middle, below). The facial expressions depict the ten emotions: ***anxious, doubtful, angry, laughing, smiling, annoyed, sad, helpless, sulking and grimacing***.

Previously at Emerj, we covered use cases wherein the company of focus highlighted the importance of variety in its dataset. For example, in a fraud detection use case, Mastercard required an array of data such as time of day, geolocation, risk assessment, and merchant information, among other data types to train their software. Mastercard claimed that the data enabled the company to offer a solution that decreases fraud, lowers operating costs, and increases revenue for its clients.

To accomplish data diversity for its solution, Clickworker asked workers on its platform to submit selfie photos from both the present and the past, thus enriching its data to account for age-related changes. Users were also required to take selfie photos at different times throughout the day, thus potentially accounting for any changes in facial expression or environment. Additionally, says Rozsenich, users skilled at labeling data to analyze and train algorithms were recruited to detect and account for changes in an individual's facial expressions.

With the crucial step of acquiring a robust dataset complete, the final act was to clean up and enrich the data where appropriate.

Clickworker then collected, collated, and screened the output data through a quality assurance process. The data was then transferred via an API connection and made available to the client via direct download using cloud-based storage.

Results

The resulting data-set provided to the client consisted of nearly ***500,000 photos*** from individuals showing their aging over a period of ten years with faces annotated. All photos were quality checked to ensure production according to specs with regards to the photo quality and diversity KPIs.

USE CASE #2

The Value of Data Quality and Diversity in Voice Recognition

Business Challenge

A client using voice recordings to identify users based on vocal patterns tasked Clickworker with collecting high-quality audio training data. To accomplish this task, Clickworker once again turned to its platform users, who submitted a vast amount of input data.

Rozsenich states that collecting a dataset capable of training NLP models to detect vocal nuance is inherently challenging. Similar to faces, voice recognition models must be able to identify subtle differences down to the individual level. Rozsenich explains,

> ***"If you think about your voice, it changes with the daily mood that you have. In the morning, you may still be a bit cranky from the night before from having a drink too much, or if you feel depressed, your voice will sound very different to someone else."***
>
> **Christian Rozsenich,** CEO of Clickworker

In a previous podcast here at Emerj, Michael Johnson, the director of research and innovation for Interactions LLC in Boston, provided his insight into the unique challenges of data collection and the implications for NLP models.

> ***"Let's say you're a major brand, and you want to be able to provide service to your customers. You want to meet them where they are, where they want to speak. You don't want to tell them what to say or force them to [use] controlled language. [The model] needs to work regardless of background noise, kids talking at the same time, accent, all of those factors."***

Because of these variations, described by Mr. Rozsenich and Mr. Johnson, any data used to train the model must therefore include data with these attributes (background noise, accent, etc.) to train the model for edge cases.

Besides variations in mood, Clickworker also had to contend with some of the obstacles described by Mr. Johnson, including background noise. To overcome these challenges, the company had to ensure the data provided by its users - and refined by its Clickworkers - was of quality, rich and diverse. We may be able to deduce some of the AI used to assist with this process; more on this below.

Turning Data into Value

To overcome the challenge of individual vocal modifications, Rozsenich states that the company required the project's ***25,000 participants*** to record three to five audio clips per session at intervals of two to three hours. The company devised this method ***"so we could get sort of a daily diary of what the voice of a certain person sounds like,"*** Rozsenich states. He says that this was a sizable enough sample with which to begin training the algorithm.

With regards to the challenges presented by device capability and environment, Rozenich states that Clickworker was prepared.

> ***"That's what we had built into our platform. We could actually simulate different situations. Capturing these different kinds of situations in the audio recordings in terms of diversity is really important to make [the data] applicable to different kinds of environments."***

Rozenich explains that its users were given a briefing that detailed the recording requirements, including the appropriate ambient noise level in which to record. Additionally, users were instructed on different types of voices to use, such as whether to whisper, shout, speak normally, etc.

Noise levels were probed by the Clickworker app when recording and appropriate filters applied to provide representative recordings. Recordings are then verified by other crowd reviewers to ensure they were produced according to specs and then annotated using ML based annotation to identify word boundaries of phrases.

Results

As a result, a set of annotated audio files, consisting of ***5,000 audio recordings*** in five different languages were delivered to the client. This data-set was then used to train and validate the client's ML algorithm. The global crowd coverage allowed the client to source this data from one vendor, rather than engaging multiple agencies, in less than four weeks time.

Concluding Takeaways

As Mr. Rozsenich alluded to at the onset, acquiring data with the traits of quality and diversity is not an easy task - and one that an enterprise may want to delegate externally. We would like to elaborate a bit more on what this process - both correctly and incorrectly done - looks like according to our experience at Emerj.

The first step lies in appropriately sourcing the data - a mostly rudimentary task for any data scientist; yet, surprisingly, one with which many enterprises struggle. Organizations will often acquire data in a way that renders mute what they are attempting to train a model to accomplish. In a previous podcast at Emerj, Daniel Hernandez, General Manager of the IBM Data and AI Business Unit elaborated on his experience with the problem of inappropriately sourced data and its effects:

> ***"Today, the majority of our customers are copying data from wherever it originates and consolidating it into a single place. The problem with that is it's expensive [and] causes all sorts of nasty ripple effects, data quality issues that if you're to remediate in one place are hard to remediate in another."***
>
> **Daniel Hernandez,** General Manager of the IBM Data and AI Business Unit

So the first step in acquiring quality, diverse data is sourcing (and storing) it correctly. This first step, as Mr. Hernandez states, is crucial for the remaining processes to work. Next, after the data has been sourced and stored appropriately, is the process of evaluating the data for quality.

Quality training data are diverse in that they are relevant, consistent, and comprehensive. The data must be relevant in that it must only include the attributes required to train the desired model. Identifying these attributes is a task best reserved for domain experts with a good understanding of the subject area. The data must also be consistent in that all attributes align to the appropriate labeling, and comprehensive in that the data has enough volume and parameters to train the model across a range of edge cases.

In the simplest terms, quality training data boils down to quality and diversity. But who is evaluating the data for these attributes? Data scientists and other experts are - from sourcing and acquisition to cleanup and deployment (or, in the case of training data, packaging.) It is in this context that Mr. Rozsenich gives perhaps his most salient advice.

When asked what advice he would impart to business leaders when it comes to assessing their data needs, Rozsenich emphasizes this point of ensuring data familiarity while cautioning on the price paid for not doing so:

> ***"Spend enough time understanding the data that you have or the data you are harvesting. Spending a lot of time in understanding the demographics [and] the technical parameters of your data is really crucial because most of the projects for AI training fail because the original data doesn't have the quality or doesn't have diversity."***
>
> **Christian Rozsenich,** CEO of Clickworker

About ClickWorker:

With more than 6,000,000 global users – known as Clickworkers – clickworker is one of the leading suppliers of paid crowdsourcing.

Clickworker offers scalable solutions in 18 languages and more than 30 target markets, including training data for AI and machine learning systems.

For standardized tasks in the areas of text production, sentiment analysis and surveys, clickworker in addition offers a self-service solution via the online marketplace: https://marketplace.clickworker.com/

Visit:
clickworker.com

Contact:
clickworker.com/contact

About Emerj Artificial Intelligence Research:

Emerj Artificial Intelligence Research is a market research and advisory company focused exclusively on the business impact of AI.

Companies that thrive in AI disruption run on more than just ideas. They leverage data and research on the AI applications delivering return in their industry today and the AI capabilities that unlock true competitive advantage into the future - and that's the focus of Emerj's research services.

Leaders in finance, government, and global industries trust Emerj to cut through the artificial intelligence hype, leverage proven best-practices, and make data-backed decisions about mission-critical priorities.

Visit:
www.emerj.com

Contact:
research@emerj.com