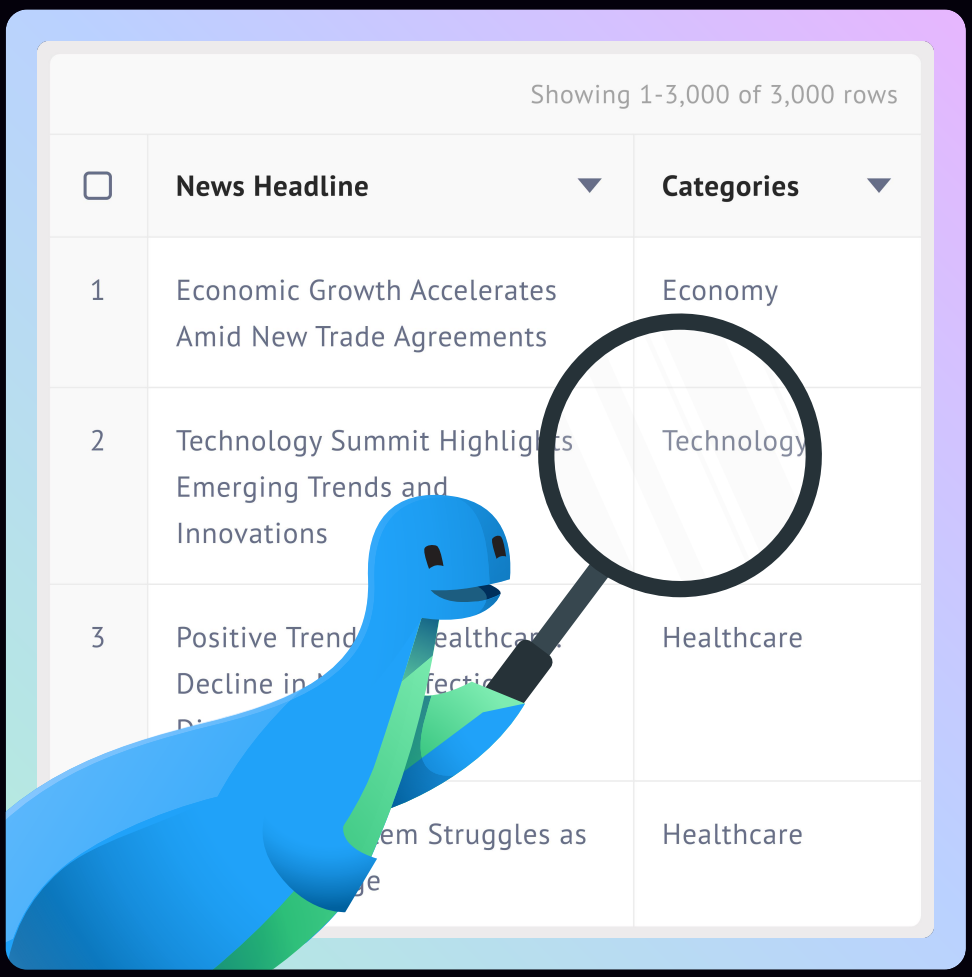



# Enhancing Data Quality: Finding and Fixing Label Errors with Datasaur

Datasaur

**Whitepaper**
June 2024

# Abstract

Building a high-quality dataset is crucial but time-consuming. While AI research often focuses on models, the importance of quality data cannot be overlooked. Label errors, in particular, can significantly impact model performance. Detecting and correcting these errors is essential for reliable datasets. Datasaur introduces label error detection, leveraging models to automatically identify and correct label errors, thus enhancing the labeling experience. Despite finding errors throughout all the data, this feature allows users to focus on a small subset of samples that need more attention, as they are the most likely errors in the dataset. Through experiments and case studies, our approach has proven to improve dataset quality and, consequently, machine learning model performance.

# Introduction

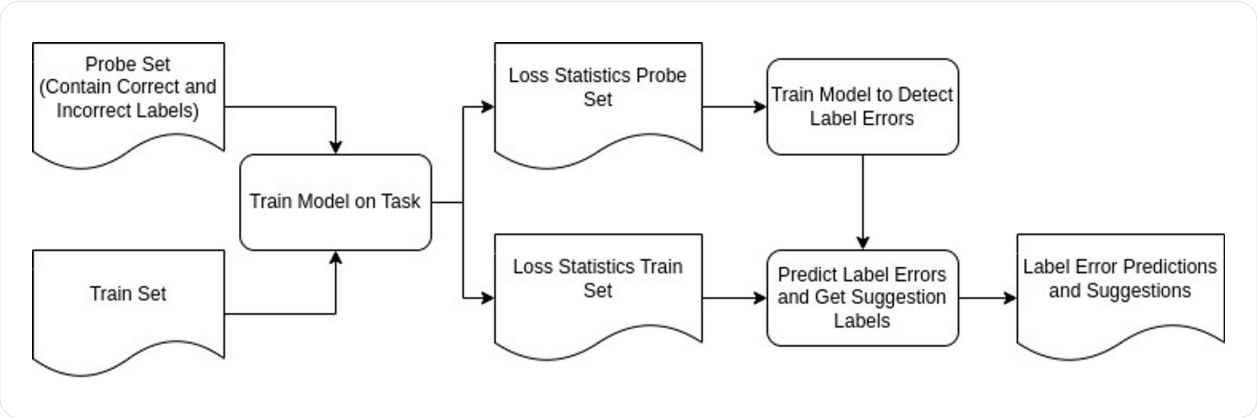
According to Andrew Ng (Ng, 2021), data preparation consumes 80% of the time, leaving the remaining 20% for actions such as model training. There are two approaches to addressing machine learning problems: model-centric and data-centric. In the model-centric approach, the focus is on refining the model while keeping the data fixed to improve accuracy. Conversely, in the data-centric approach, the emphasis shifts to improving the data while keeping the model constant. Despite the availability of advanced off-the-shelf models and benchmarks, achieving significant improvements in accuracy is challenging, as differences among models are often minimal (Ng, 2021).

What if we attempted to solve machine learning tasks using a data-centric paradigm? By improving the quality of the data through a model-centric approach, we can enhance the accuracy of the model without modifying its code. Andrew emphasizes the criticality of labeling consistency, data cleansing, and error correction within our datasets (Press, 2023). To systematically support these needs, a labeling pipeline is required. The Datasaur app offers a labeling system that promotes consistency by enabling label consensus to measure agreements among labelers and corrects errors through label error detection. By integrating traditional review processes with machine learning models, the labeling process can be executed effectively and efficiently.

Datasaur is excited to introduce Label Error Detection, a groundbreaking feature powered by Metadata Archaeology (Siddiqui et al., 2022), proposed to identify label errors by monitoring model loss throughout the training process for each individual example. Loss statistics for each data point serve as indicators for metadata such as erroneous, in-distribution, out-of-distribution, majority group, minority group, and more. This feature is designed to enhance your labeling experience by automatically identifying and correcting issues within your dataset.

# Approach

Metadata is data about data. It provides information about properties such as format, type, data history, and also the quality of the data itself. Metadata is essential for the analysis and audit of datasets for instance we can group the data to the majority or minority class, falls within or outside distribution, or has been labeled correctly. Understanding metadata simplifies the assessment of dataset quality and is crucial for developing effective labeling strategies and training models. For example, if we identify a subset of data categorized as out-of-distribution, we may consider excluding it from the training data to prevent any negative impact on the model's performance.




Picture 1. Metadata archaeology to detect label errors. This method involves two models. The first model captures loss statistics during training on a specific task, while the second model uses these statistics loss as features to learn the training dynamics and detect label errors.

Datasaur integrates Metadata Archaeology to streamline the labeling process by identifying and correcting label errors. This method involves curating two subsets, known as probe suites: one representing correct labels and the other incorrect labels. Analyzing the training dynamics of the dataset alongside these curated subsets enables us to infer the metadata associated with each data point, allowing us to determine whether its label is correct or incorrect.

Preparing curated subsets of data can be done manually or automatically. In this system, we prefer creating curated probe suites through simple transformations and generating only a few annotated probe examples, as human annotations can be quite expensive

Metadata archaeology via probe dynamics leverages the loss statistics throughout the training process on each individual example. The system records the model's loss and learns the differences in learning patterns for each probe category. This loss statistics could have various implications for defining metadata such as outliers or inconsistencies in labels within our dataset, but here we focus on correct and incorrect labels.

# Case Study

We conducted experiments utilizing metadata-archaeology via probe-dynamic on classification tasks. Using publicly available datasets such as Agnews for news articles, IMDB for sentiment classification of movie reviews, yelp_polarity containing Yelp reviews for sentiment analysis, and Dbpedia_14 for classifying Wikipedia content. To simulate the data error during our experiments, we sampled the data and altered the labels since there were no specific datasets available to address this task. We assumed each dataset contained a 10% margin of error. This allowed us to measure how well our integrated label error detection identified errors using the f1-score metric.

| No | Dataset | Number of Class | Number of Sample | Number of Errors | F1-Score |
|---|---|---|---|---|---|
| 1 | ag_news | 4 | 12.000 | 1.200 | 72.4% |
| 2 | IMDB | 2 | 10.000 | 1.000 | 47.3% |
| 3 | yelp_polarity | 2 | 28.000 | 2.800 | 65.1% |
| 4 | dbpedia_14 | 14 | 28.000 | 2.800 | 95.3% |

Table 1: Performance of label error detection across five datasets

Our system shows the ability to identify label errors in four different datasets, helping users reduce errors and improve dataset quality. The F1-Score metric is suitable for measuring our system's performance because it considers the small subset of data where errors often occur (imbalance case). Reviewing a large amount of data is not easy or efficient. Our system enhances the labeling experience by highlighting the most important samples for review, making the process more efficient.

| No | Dataset | 50 samples | 100 samples | 500 samples | 1000 samples | 5000 samples |
|---|---|---|---|---|---|---|
| 1 | ag_news | 22.2% | 43.2% | 61% | 59% | 72.5% |
| 2 | IMDB | 25% | 21.9% | 38.4% | 38% | 44.8% |
| 3 | yelp_polarity | 25% | 28% | 40% | 48% | 60% |
| 4 | dbpedia_14 | 34.5% | 23.5% | 73.9% | 71% | 91.5% |

Table 2: The performance of label error detection (LED) using the F1-Score metric is evaluated across datasets containing varying numbers of samples, each with 10% label errors.

Based on experiments on table 2, It's observed that more data leads to better performance of label error detection (LED) in identifying issues within datasets. Implementing label error detection can be particularly helpful for users with large amounts of data to review. Instead of reviewing all data points, which can be time-consuming, LED suggests focusing on a small subset of data with the highest possibility of errors.

| No | Dataset | Number of Sample | Total error found | Thres > 0.6 | Thres > 0.7 | Thres > 0.8 | Thres > 0.9 | Thres > 0.95 |
|---|---|---|---|---|---|---|---|---|
| 1 | ag_news | 7.600 | 531 (7%) | 494 (6.5%) | 451 (5.9%) | 280 (3.7%) | 154 (2%) | 56 (0.7%) |
| 2 | IMDB | 25.000 | 4.532 (18.1%) | 3.686 (14.7%) | 1.728 (6.9%) | 715 (2.9%) | 273 (1.1%) | 163 (0.7%) |
| 3 | yelp_polarity | 38.000 | 4.487 (11.8%) | 3.557 (9.4%) | 2.878 (7.6%) | 2.150 (5.7%) | 1.198 (3.2%) | 805 (2%) |
| 4 | dbpedia_14 | 70.000 | 1.179 (1.7%) | 946 (1.7%) | 946 (1.7%) | 226 (0.3%) | 226 (0.3%) | 226 (0.3%) |

Table 3: Label Error Detection is performed on public datasets using a test set.

According to Table 3, we aim to determine the number of data points with potential label errors that require review. By selecting the most probable label issues, we make it easier to identify label errors in datasets rather than reviewing all samples. Based on our empirical study conducted through experiments across four datasets, we recommend that users perform label error detection on hundreds or thousands of samples to experience its impact. Since label error detection employs a machine learning approach, we understand that more data typically results in improved performance in identifying label issues.

Label error detection also allows users to adjust the error possibility, with higher numbers indicating a higher probability of errors. This feature caters to users with different needs; for instance, those uncertain about their label quality may opt for a lower threshold, while those confident in their label quality may choose a threshold greater than 0.9, potentially reducing the number of data points requiring review to as little as 1%.

| No | Dataset | F1-Score of Original data | F1-Score with AutoCorrection |
|---|---|---|---|
| 1 | ag_news | 90% | 93.47% |
| 2 | IMDB | 90% | 90.59% |
| 3 | yelp_polarity | 90% | 93.49% |
| 4 | dbpedia_14 | 90% | 98.68% |

Table 4: Improvement in Data Quality through Label Correction

Table 4 displays our system's capability to improve dataset quality of 4 datasets through label correction. The provided label suggestions serve as additional information, highlighting potential challenges for the model in selecting the best label when confidence scores between two labels are close together. However, it is important to include a disclaimer stating that auto-correction should not be relied upon solely; human review is always preferable in this scenario.

# Finding Errors in Publicly Available Datasets

We also conducted label error detection to find errors in the labels of public test datasets such as AG News, IMDB, and Yelp Polarity. We sampled data that appears to have inconsistent answers based on the given labels.

| No | Text | Given label | Datasaur suggestions |
|---|---|---|---|
| 1 | Google Lowers Its IPO Price Range SAN JOSE, Calif. - In a sign that Google Inc.'s initial public offering isn't as popular as expected, the company lowered its estimated price range to between \$85 and \$95 per share, down from the earlier prediction of \$108 and \$135 per share... | world | business |
| 2 | Indonesian diplomats asked to help improve RI #39;s bad image JAKARTA (Antara): President Susilo Yudhoyono asked Indonesian diplomats on Monday to help the government improve Indonesia #39;s bad image. | business | world |
| 3 | Davenport Advances at U.S. Open NEW YORK - Lindsay Davenport's summer of success stayed on course Thursday when the fifth-seeded former U.S. Open champion defeated Arantxa Parra Santonja 6-4, 6-2 and advanced to the third round of the season's final Grand Slam event... | world | sports |
| 4 | Bush Scraps Most U.S. Sanctions on Libya (Reuters) Reuters - President Bush on Monday formally\ended the U.S. trade embargo on Libya to reward it for giving\up weapons of mass destruction but left in place U.S.\terrorism-related sanctions. | science & technology | world |
| 5 | Philippines mourns dead in Russian school siege The Philippines Saturday expressed quot;deepest sympathy quot; to the families of the dead in the Russian school siege on Friday, in which 322 people were killed when Russian troops stormed | science & technology | world |

Table 5: Sample of label errors with our system using AG News dataset

| No | Text | Given label | Datasaur suggestions |
|---|---|---|---|
| 1 | Just watched on UbuWeb this early experimental short film directed by William Vance and Orson Welles. Yes, you read that right, Orson Welles! Years before he gained fame for radio's "The War of the Worlds" and his feature debut Citizen Kane, Welles was a 19-year-old just finding his muse. Besides Vance and Welles, another player here was one Virginia Nicholson, who would become Orson's first wife. She plays a woman who keeps sitting on something that rocks back and forth courtesy of an African-American servant (Paul Edgerton in blackface). During this time a man (Welles) keeps passing her by (courtesy of the scene constantly repeating). I won't reveal any more except to say how interesting the silent images were as they jump-cut constantly. That's not to say this was any good but it was fascinating to watch even with the guitar score (by Larry Morotta) added in the 2005 print I watched. Worth a look for Welles enthusiasts and anyone with a taste of the avant-garde. | negative | positive |
| 2 | This is definitely one of the best Kung fu movies in the history of Cinema. The screenplay is really well done (which is not often the case for this type of movies) and you can see that Chuck (in one of his first role)is a great actor. The final fight with the sherif deputy in the bullring is a masterpiece! | negative | positive |
| 3 | Nice movie and Nicholle Tom does a fantastic job playing the "guy in the girl's body", she really does it well. A sort of teen version of many other movies, but well done. Well casted, from "Matt" to "Matt2". | negative | positive |
| 4 | I'm going to make this short and sweet. It's not surprising that you had no use for this film. This is a story about the power, beauty and possibilities inherent in a meaningful education. Based on your pathetically composed comments I can see that your own education has been woefully neglected... or worse... completely wasted. Your comments are those of a truly ignorant person. I would advise you to do something about this condition... but in your case I feel it's probably too late. My hope is that you yourself don't intend to go into the teaching profession ( especially in Film Studies) because you could only do damage. Oh... one last bit of advice. In the future, if you intend to write more opinion pieces, you should really proofread your work. It will make people take you more seriously. | positive | negative |

| | | | |
|---|---|---|---|
| 4 | I'm going to make this short and sweet. It's not surprising that you had no use for this film. This is a story about the power, beauty and possibilities inherent in a meaningful education. Based on your pathetically composed comments I can see that your own education has been woefully neglected... or worse... completely wasted. Your comments are those of a truly ignorant person. I would advise you to do something about this condition... but in your case I feel it's probably too late. My hope is that you yourself don't intend to go into the teaching profession ( especially in Film Studies) because you could only do damage. Oh... one last bit of advice. In the future, if you intend to write more opinion pieces, you should really proofread your work. It will make people take you more seriously. | positive | negative |

Table 6: Sample of label errors with our system using Yelp Polarity dataset

| No | Text | Given label | Datasaur suggestions |
|---|---|---|---|
| 1 | not the best | positive | negative |
| 2 | Its ok. Its an airport, and a stereotypical one at that. It has long lines and a lot of traffic, expensive chain restaurants, and shops. Nothing special. Customer Service is passable and they do a good job as expected, however it is a typical airport. No muss or fuss. | positive | negative |
| 3 | I would recommend you the potatoes soup or clam chowder they are awesome... services is ok, kind of slow during lunch time | negative | positive |
| 4 | Food okay, great tv watching and service can use some help. Overall not bad and would visit again but more with sports watching and not just dinner. | negative | positive |
| 5 | Says they deliver on here... Wrong & wrong again & should not be checked! I like Applebee's & thought the delivery was something new for the Pittsburgh area... Don't know if this is something Yelp does or something someone checked off... But not cool! | positive | negative |

Table 7: Sample of label errors with our system using IMDB dataset

# Conclusion

Label error detection provides a practical solution for identifying issues in datasets using metadata archaeology through training dynamics. This feature streamlines the reviewing process, enabling users to focus on a subset of data flagged as label errors.

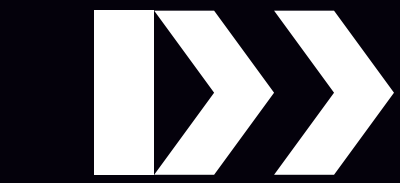
As a result, reviewers can adjust the error probability to focus on the most probable errors, thereby reducing their reviewing costs by up to 99% compared to manually reviewing all data points. We believe that curated data, powered by machine learning algorithms, can enhance the quality of data and improve the performance of machine learning models. This research contributes to applied machine learning in labeling platforms, enhancing the overall labeling experience.

# References


- Siddiqui, S. A., Rajkumar, N., Maharaj, T., Krueger, D., & Hooker, S. (2022). *Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics* (arXiv:2209.10015). arXiv. http://arxiv.org/abs/2209.10015

- Ng, A. (2021, March 25). *A chat with andrew on MLOps: From model-centric to data-centric AI*. YouTube. https://www.youtube.com/watch?v=06-AZXmwHjo

- Press, G. (2021, June 16). *Andrew Ng launches a campaign for data-centric AI*. Forbes. https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=236cc00074f5



Datasaur