Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 56–65
Marseille, 20 June 2022

Annotator Response Distributions as a Sampling Frame

Christopher M. Homan, Tharindu Cyril Weerasooriya, Lora Aroyo, Chris Welty
Rochester Institute of Technology, Google
USA
cmh@cs.rit.edu, {cyriltcw, l.m.aroyo, cawelty}@gmail.com

Abstract

Annotator disagreement is often dismissed as noise or the result of poor annotation process quality. Others have argued that it can be meaningful. But lacking a rigorous statistical foundation, the analysis of disagreement patterns can resemble a high-tech form of tea-leaf-reading. We contribute a framework for analyzing the variation of per-item annotator response distributions to data for humans-in-the-loop machine learning. We provide visualizations for, and use the framework to analyze the variance in, a crowdsourced dataset of hard-to-classify examples of the OpenImages archive.

Keywords: preserving disagreement, statistical methods, empirical study

1. Introduction

With a market expected to hit $1.2 billion by 2023, human annotation accounts for 80% of the time spent building A.I. technology. (Metz, 2019). Whether obtained by a small team of experts, or an anonymous pool of crowdworkers, it is generally considered good practice to obtain responses from multiple annotators for each example in a dataset, for the reason that human annotators are unreliable and annotation tasks are ambiguous. And so disagreement is seen as a sign of something to be corrected. Put more formally, machine learning problems are probability distributions over a joint (example, response) space $\mathcal{X} \times \mathcal{Y}$ (Shalev-Shwartz and Ben-David, 2014). Usually, the distribution over $\mathcal{Y}$ has a Bayesian interpretation, where $\mathrm{P}(y \mid x)$ is seen as uncertainty over the response.

An alternate view is that disagreement is meaningful and may be the result of differences in annotator values, beliefs, or values that carries meaningful signals (Aroyo and Welty, 2015; Liu et al., 2019; Akhtar et al., 2019; Klenner et al., 2020; Weerasooriya et al., 2020; Davani et al., 2022; Basile, 2020). We are particularly interested in crowdsourced settings, where there are typically more annotators per example than with expert annotations. Taking a strictly frequentist approach, we interpret $\mathrm{P}(y \mid x)$ as the likelihood of drawing an annotator who responds to example x with y. We are thus interested in asking *How confident are we that* $\mathrm{P}(y \mid x)$ *represents the ground truth distribution of annotator responses?*.

We apply hypothesis tests via bootstrap sampling (Efron, 1992) to explore this question on a dataset that is particularly rich in annotatator disagreement. A major design decision in this case is which test statistic to use. If we were measuring machine performance, we could use any number of standard evaluation measures, such as accuracy or precision. But here, we need a statistic that can measure the difference in two probability distributions. Many exist, such as KL-divergence and Wasserstein distance. However, these measures do not take into account that our distributions are merely samples. We argue that the likelihood function of the hypothesized sampling frame is the best test statistic in this case.

In this paper, we contribute a framework for analyzing the variance of annotator responses in machine learning training data when the goal is to preserve diversity in annotator responses by treating them as a sample from an underlying pool of respondents. We introduce two variants of bootstrap sampling tailored to this setting that are more efficient and/or less sensitive to sparse data than true bootstrapping. We explore the use of the log-likelihood as a statistic for hypothesis testing in exploratory analyses of response distribution data. And we apply this framework to an empirical study of a data set rich in annotator disagreement.

2. Related Work

Although not as commonly used as in other scientific fields, hypothesis tests has a long history in machine learning (Mitchell, 1997).

Dietterich (Dietterich, 1998) provides a taxonomy of use cases for hypothesis testing on machine learning problems. He focuses on one particular case: that of choosing between two learning algorithms A and B with a small amount ($n \approx 300$) of data. He defines the p-value to be the probability that A's error is less than B's by at least the observed error difference $\delta(\mathbf{x})$, where $\mathbf{x}$ is a sample from the test population. assuming as the null hypothesis H_0 that A and B have equal error rates in the population from which $\mathbf{x}$ was sampled. Formally, this is denoted $p(\delta(\mathbf{x}^*) > \delta(\mathbf{x})|H_0)$, where $\mathbf{x}^*$ is a population sample of size n drawn according to H_0. Thus, in contrast to our paper, he is interested in paired hypothesis tests, as is frequently the case in machine learning.

He compares five different approximations of the p-value on experiments where A and B are simulated and by design have the same error rate, though their responses differ on specific items. He repeats these

experiments using two actual (i.e., nonsimulated) machine learning algorithms, where one is "hobbled" to have exactly the same error rate as the other. In this setting he tests the approximations' resistance to Type I errors, as well as their statistical power in the event that the two algorithms do have different error rates.

Berg-Kirkpatrick et al. (Berg-Kirkpatrick et al., 2012) perform an empirical investigation of hypothesis testing across a seven natural language processing (NLP) problems. They survey prior work on these problems where the systems were available for evaluation, and study the relationship between metric gain, $\delta(\mathbf{x})$, statistical significance, and p-values. They argue that the best approach is to bootstrap from the input sample and then consider $p(\delta(\mathbf{x}^*) > 2\delta(\mathbf{x})|H_0)$. Søgaard et al. (Søgaard et al., 2014) study the practical impact of various estimators on p-values.

Reidsma and Carletta (Reidsma and Carletta, 2008) explore the relationship between interrater reliability and machine learning performance. They show that high reliability scores ($> .8$) predict good machine learning performance *as long as noise is unbiased*. If noise is biased, the machine learning algorithm may learn the bias pattern and overfit.

Szymański and Gorman (Szymański and Gorman, 2020) apply a Bayesian framework due to (Corani et al., 2017) to evaluate the performance of English part-of-speech taggers. Rather than p-values based on H_0, their framework estimates the likelihood that system A outperforms system B, using k-fold cross evaluation (across multiple datasets). Zhang et al. (Zhang et al., 2004) use bootstrapping to construct confidence intervals for BLEU scores.

Welty et al. (Welty et al., 2019) study the problem of measuring AI systems from the perspective of *metrology*, the science of measurement and its application. They demonstrate these principles on WordSim (WS353),[1] a crowd-powered dataset for word similarity. They show that the dataset can be *instrumentalized* by describing procedures for (1) collecting the human or crowd data (2) using this data to evaluate the performance of an AI system. The introduce a number of key concepts in metrology and show how they apply in this context. For instance, the *principle of measurement* translates into understanding the limits and opportunities of the measurement frame and how the measurement procedure works, and *indication* translates into the itemwise statistics gathered from asking multiple annotators about the same question. A crucial element of metrology is the recognition that ground truth is fundamentally unknowable, and that one must test and assess the accuracy of any instrument used to measure performance.

[1] https://aclweb.org/aclwiki/WS353ilarity-353_Test_Collection_(State_of_theart)

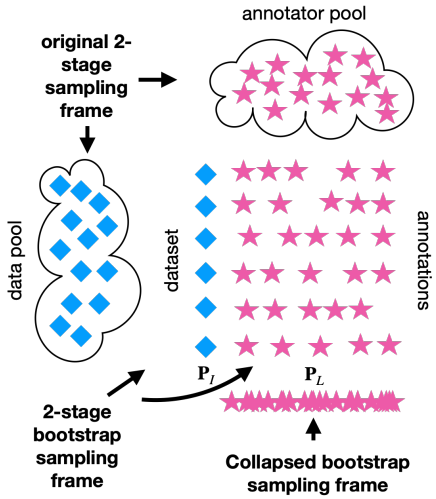


Figure 1: Bootstrapping is a stochastic method for analyzing variance in samples. It treats the sample as an estimate of the underlying (original) sampling frame, and then repeatedly samples with replacement from the empirical sample, obtaining a sample of samples. Annotator sampling is itself a two-stage process, where the empirical sample consists of first drawing from a set of data items (in our case image/label pairs from the Open Images Dataset) and, for each item, sampling from a pool of annotators. However, when the space of annotator responses is relatively simple, we can marginalize over the data items to create a collapsed, one-stage bootstrap sampling frame.

3. Annotator sampling

Here, we describe three variants of bootstrapping that we explore in this paper. We adapt notation from (Efron, 1992). Suppose we have a set of m data items $\mathrm{x} = (x_1, \ldots, x_m)$, sampled from some domain F_I. For each item i, we also have a sample y_i of r annotator responses, where each response comes from a discrete domain of q options, indexed by l. There are multiple ways to represent $\hat{y}_i$. For each response l, we can count the number of annotators what respond with l, which we denote $\hat{y}_{i,l}$. Or we can indicate the response that annotator j provides, which we denote $y_{i,j}$. Note that we use y with and without the $\hat{\cdot}$ in part to distinguish these two representations, but also to stress that $\hat{y}_{i,l}$ is not necessarily representative of the underlying population's value for the number of l responses (assuming the underlying population of annotators is much larger than the number of responses in y_i, it is most certainly a much larger number), where $y_{i,j}$ is in fact annotator j's response to item i. We can extend this latter representation to the set of all annotations as a matrix, where the data examples are aligned along the vertical axis and the responses along the horizontal. See Figure 1.

Finally, we can represent y_i as a distribution $\hat{F}_{y_i}$.

Bootstrapping (Efron, 1992) is a stochastic method for estimating the variance of a *test statistic* ϕ from any *empirical sample* $\mathbf{x}$. It constructs a sample of B samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, where each of these latter samples is the same size as the empirical sample and is drawn with replacement from the empirical sample, effectively using the empirical sample as an estimate $\hat{F}_I$ of the original sampling frame F_I. Thus, in our setting, each bootstrap sample $\mathrm{x}^{*j} = (x_1^{*j}, x_2^{*j}, \ldots, x_m^{*j})$ consists of m items sampled with replacement from $\mathrm{x} = (x_1, \ldots, x_m)$. In this way it can account for the impact of sample size on the variance of any test statistic, though if the empirical sample is too small to be representative of the original sampling frame the method can be ineffective.

In the past, when bootstrapping was used to analyze variance in machine learning datasets (Mitchell, 1997; Dietterich, 1998; Zhang et al., 2004; Berg-Kirkpatrick et al., 2012; Søgaard et al., 2014), it was performed over the items only, i.e., in the vertical direction only according to the matrix-style representation shown in Figure 1. In the parlance of our notation, each item x_i^{*j} in each bootstrap sample $\mathbf{x}^{*j}$ is associated with the same label y_i^{*j} as its corresponding empirical item. Of course, in most past settings, y_i represented a single response value, as all annotator disagreement was typically resolved before the data was used, and so this vertical-only approach made perfect sense.

As a baseline, we adapt this strategy to our case, i.e., we associate each item x_i^{*j} in each bootstrap sample $\mathbf{x}^{*j}$ with the empirical distribution y_i^{*j} associated with the corresponding empirical item. We call this vertical-only baseline process a *naive bootstrap*.

However, in case of annotator modeling, where we care about the ground truth distribution of annotator responses, the empirical sample is really the result of two-stage process. See Figure 1. First, choose a data item i in the vertical direction, then choose r annotators in the horizontal direction to annotate it.[2]

In many datasets the number of annotators r varies from item to item. But (as in the case of the data we analyze here) if r is the same for each item, then the number of possible response distributions is $\binom{q+r-1}{r-1}$, and when this number is sufficiently small we can simplify boostrapping over this two-stage process by precomputing the horizontal bootstrap and marginalizing over the examples i. Thus, we construct a distribution $\hat{F}_{q,r}$ over all annotator response distributions y^* of size

[2]This is a simplification of how annotation works in practice. Typically, annotators are not chosen independently for each item, as we assume here. However, for large datasets, as long as the number of items any one annotator sees is small—as is often the case for crowdsourced annotations—we do not believe dependencies between annotators have a significant impact on the analysis described here, although this is certainly a topic worthy of future research.

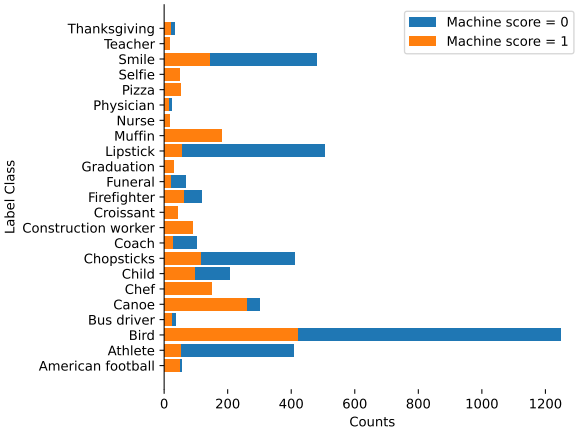


Figure 2: Counts of image/label pairs by machine score.

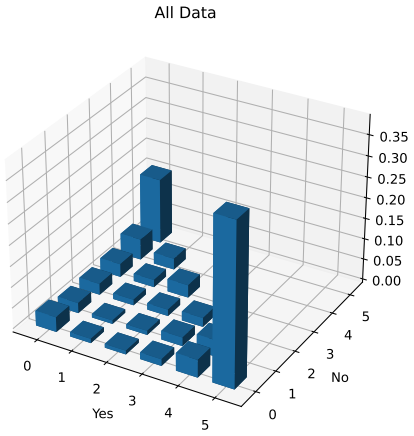
r:

$$\hat{F}_{q,r}(y^*) = \sum_{i=1}^{m} \hat{F}_{y_i,r}(y^*)\hat{F}_I(x_i) = \frac{1}{m}\sum_{i=1}^{m} \hat{F}_{y_i,r}(y^*), \tag{1}$$

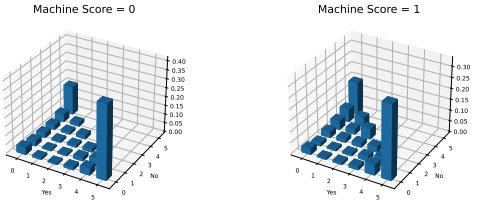
where $\hat{F}_I(x_i) = \frac{1}{m}$ per the rules of bootstrap sampling, and $\hat{F}_{y_i,r}(y^*)$ is the likelihood of drawing the distribution y^* by drawing (with replacement) a sample of size r from $\hat{F}_{y_i}$. We call this approach *collapsed bootstrapping*. Collapsing can greatly speed up the sampling process by eliminating one stage of sampling. Moreover, it removes some of the stochasticity from the process. This, in turn, means that a smaller bootstrap sample is needed.

Finally, many of the annotator response distributions themselves may have no mass on some of the responses (e.g., cases where all five annotators agree on a single response). Therefore, it may make sense to add smoothing to the collapsed distribution. We use Laplace smoothing, with $\alpha = 1$, which assumes a uniform prior over all choices, and we apply this to both stages (i.e., to each $\hat{F}_{y_i,r}$ and to $\hat{F}_{q,r}$ in Equation 1). We call this *smoothed bootstrapping*.

Beyond the sampling process itself, bootstrapping is often used for *hypothesis testing*. This involves choosing test statistics and hypotheses. The mean of some quantity of interest is by far the most common test statistic used. But when the data under consideration (representing the sampling frame) is categorical, or if we are interested qualitatively in the shape of the distribution, KL-divergence or Wasserstein distance might be more appropriate choices. The best statistic and hypotheses to use depends what one is trying to learn from the test. So let us first introduce the dataset we are analyzing, and some of the questions we seek to answer, before considering this question further.



(a) Entire dataset.



(b) Only item-annotator pairs with $Machine_i = 0$

(c) Only item-annotator pairs with $Machine_i = 1$

Figure 3: Histograms of annotator response distributions. Each image/label is indicated by the number of *Yes* and *No* annotator responses. The number of *Don't know* responses can be calculated from the number of *Yes* and *No* annotator responses and so is not shown. For instance, the $(0, 0)$ corner represents the number of images where all five responses were *Don't know*. The two large peaks in the corners are the item-label pairs on which all annotators agreed on *No* (respectively, *Yes*). The much smaller peak in the left-hand corner are the item-label pairs on which all annotators agreed on the *Don't know*. Note that there appears to be more disagreement among the image/pairs with $Machine_i = 1$.

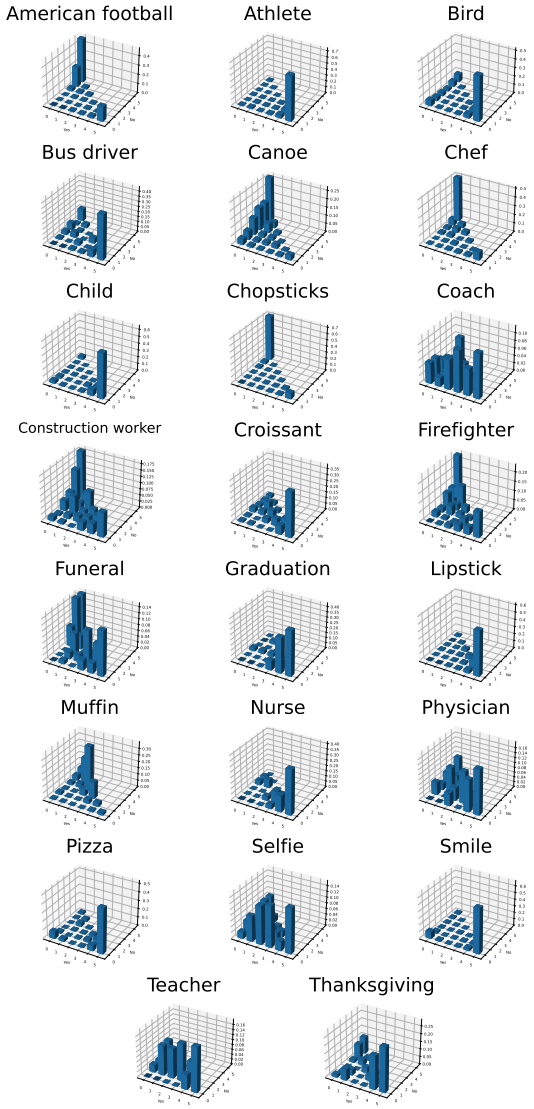


Figure 4: Annotator response distributions by label class.

4. Data

The CATS4ML (Crowdsourcing Adverse Test Sets for Machine Learning) Data Challenge[3] asked participants to find machine learning *blind spots*, i.e., data instances that humans can easily classify, but on which machine learning algorithms fail.

The data consists of $6,393$ examples of image/label pairs from the Open Images Dataset (OID). The labels in these image/label pairs were selected from among 23 label classes, which were sampled from 30K classes available in the OID. Note that "label" often refers to the annotator responses y_i. Here and throughout this paper, we use "label" only to refer to the label class, which is part of the input, and not the annotator responses.

As for the responses, for each image in each image/label pair, five annotators were asked whether the label matched the image. Each image/label pair i has a distribution F_{y_i} over $q = 3$ annotator response choices: Yes, No, $\neg Know$, which indicate the number of annotators who respond *Yes*, *No*, or *Don't know*, respectively. There is also $\text{Machine}_i \in \{0, 1\}$, a machine response, chosen by randomly sampling the output from two machine-based classifiers (variants of the InceptionV2-based classification that are internal to Google). These human and machine responses were used to adjudicate the submissions to the contest. Figure 2 shows the distribution of images/pairs in the dataset by machine score.

Since there are only three possible label responses, the space of *annotator response distributions* forms a 2-simplex (or triangle), where, since each image/label

[3] https://github.com/google-research-datasets/cats4ml-2021-dataset. See also https://cats4ml.humancomputation.com/.

Annotator rating		Machine rating 0	1	total
Plurality	Yes	2552	1024	3576
	No	1578	853	2431
Majority	Yes	2423	964	3387
	No	1284	729	2013
≥ 4	Yes	2171	840	3011
	No	998	502	1500
Unanimous	Yes	1820	667	2487
	No	708	298	1000

(a) Distribution of *Yes* and *No* annotator responses based on various disagreement resolution/exclusion policies: only those where the number of yes (respectively, no) responses exceeds no (respectively, yes), those with a majority of yes vs. no votes, those with at least four votes in agreement, and those with unanimous agreement.

Annotator rating		Machine rating 0	1	total
Plurality	Yes	2479	1119	3599
	No	1640	740	2380
Majority	Yes	2341	1057	3398
	No	1391	628	2018
≥ 4	Yes	2094	945	3040
	No	1096	495	1591
Unanimous	Yes	1855	837	2692
	No	836	377	1213

(b) Estimated number of data items by user and machine response according to the collapsed boostrap frame.

Annotator rating		Machine rating 0	1	total
Plurality	Yes	2490	1124	3614
	No	1636	739	2375
Majority	Yes	2307	1042	3349
	No	1357	613	1970
≥ 4	Yes	1968	888	2856
	No	1012	457	1469
Unanimous	Yes	1287	581	1868
	No	584	264	848

(c) Estimated number of data items by user and machine response according to smoothed ($\alpha = 1$) bootstrap frame.

Table 1: According to various bootstrap methods, the distribution of *Yes* and *No* annotator responses based on various disagreement resolution/exclusion policies: only those where the number of yes (respectively, no) responses exceeds no (respectively, yes), those with a majority of yes vs. no votes, those with at least four votes in agreement, and those with unanimous agreement.

pair i has exactly five annotator responses, i.e., $y_{i,Yes} + y_{i,No} + y_{i,\neg Know} = 5$, the vertices of the triangle represent unanimous responses (i.e., EITHER $y_{i,Yes} = 5$ OR $y_{i,No} = 5$ OR $y_{i,\neg Know} = 5$, and the remaining response choices equal to zero), and the edges and interior space represent responses that have at least some level of annotator disagreement. It is a discrete space of cardinality 21 and so it is easy to precompute the bootstrapping, as shown in Equation (1).

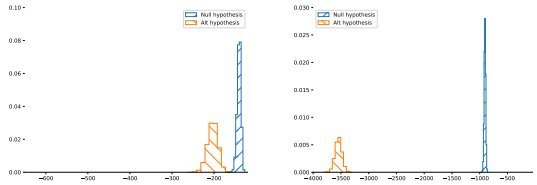
The three-option response schema used in this dataset lends itself very well to visualization. Figure 3 shows histograms in this triangle-like structure of annotator response distributions over, respectively, the entire dataset, just those image/label pairs with $\mathrm{Machine}_i = 0$, and just those with $\mathrm{Machine}_i = 1$, respectively. The differences between the three are very small, though there appears to be slightly more disagreement among the pairs with machine score 1, suggesting that the CATS4ML contestants had mixed (though reasonable given the sparsity of blind spot data) success against the reference machine responses.

Figure 4 shows these same distributions by label class. Here, in contrast to Figure 3, there appear to be significant patterns. For instance, in the *Muffin* label class there is substantial annotator disagreement among annotators between *Yes* and *No*, with very few annotators responding *Don't Know*. This may be because muffins are only well-known in the US, Canada, and Great Britain, and in the US and Canada they are sweet snacks resembling cupcakes, but in Great Britain they are flat, savory rounds of bread (known as 'English Muffins' in the US and Canada). And so in this case disagreement is not the result of a poorly formed question, and it is not even "ambiguous" in the sense that a single annotator would necessarily recognize that there are multiple interpretations.

In short, there are two obvious ways to partition the data: by the machine score used to adjudicate the CATS4ML contest, and by the label classes. The hypothesis tests we consider in this paper will help us determine whether the patterns of annotator responses seen in Figures 3 and 4 are significant.

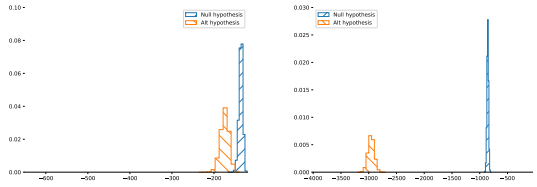
But before we get to hypothesis testing, one reason why annotator disagreement is sometimes questioned as a useful signal is because the tasks for which machine classifiers are trained often require discrete decisions (Gordon et al., 2022). But even then, the presence of disagreement requires some sort of resolution process, and the choice of a particular resolution strategy can lead to bias.

Table 1a shows how several common strategies for resolving annotator disagreement affects the distribution of the responses over examples, after resolution over the empirical annotator response distributions. Table 1b (respectively, Table 1c) shows what happens when we use collapsed (respectively, smoothed) bootstrapped frames instead (and taking the expected counts of the image/label pairs, rounded to the nearest whole number, given the sample sample size as the original dataset). The differences between the three sets are very small when the plurality response is used. This is in keeping with conventional wisdom that the number of annotators need not be very large if plurality is used to resolve disagreement (Snow et al., 2008).



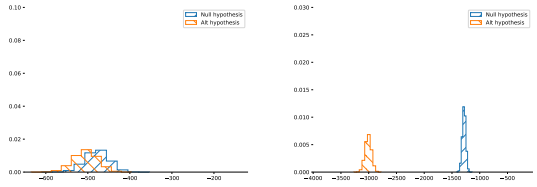
(a) Naive bootstrap ($p = 3 \times 10^{-5}$).

(e) Naive bootstrap ($p < 10^{-5}$).



(b) Collapsed bootstrap ($p = .003$).

(f) Collapsed bootstrap ($p < 10^{-5}$).



(c) Smoothed bootstrap ($p = .3$).

(g) Smoothed bootstrap ($p < 10^{-5}$).

(d) The alternative hypothesis is that the data associated with each **machine score** was generated from a distinct distribution.

(h) The alternative hypothesis is that the data associated with each **label class** came from a distinct distribution.

Figure 5: Bootstrap samples where the test statistic is *the log-likelihood of annotator response distributions under the null hypothesis*, with the null hypothesis is that the data was generated from a single distribution and the alternative hypothesis that the data associated with each **machine score** (left) **label class** (right) was generated from a distinct distribution.

However, the differences between the samples became increasingly stark as the aggregation methods become stricter.

5. Tests

We now construct tests for whether the differences observed in annotator response distributions between the data with machine scores of zero versus one, as shown in Figure 3, or with different label classes, as shown in Figure 4, are significant. For any partitioning of the dataset $D = D_1 \cup \cdots \cup D_s$ (where the partitioning might represent the different machine scores or the various label classes), let the *null hypothesis* be that the annotator response distributions y_i were sampled from the same underlying distribution F_D, as estimated by the bootstrap sampling frame over all the label distributions D. In our dataset, for naive bootstrapping this is the distribution shown in Figure 3a.

This is a very strong null hypothesis. It is much more common to define the null hypothesis in terms of a test statistic and not worry about the underlying distributions. This is because, when the null hypothesis is rejected, such weaker hypotheses tend to confer a more positive view of the test statistic, and often it is the test statistic that is of primary interest, because it is a measure of performance. But our motivation here is not to evaluate performance; rather, it is exploratory in nature. And so we are simply interested in whether the differences in the distributions we observed are meaningful. The downside to this approach is that if we reject the null hypothesis, we can only conclude that the differences observed are significant; we cannot reasonably conclude anything positive about the nature of the distributions.

As our test statistic, we use the log-likelihood of the null hypothesis:

$$\log F_D(D_1^*) + \log F_D(D_2^*) + \cdots + \log F_D(D_s^*) \tag{2}$$

Where $D_1^*, \ldots, D_s^*$ are samples of each partition under the null hypotheses, i.e., they are samples of the bootstrap frame F_D.

As for computing the p-value, we could, for each bootstrap sample, compare the value of Equation (2) to the log-likelihood of the original sample $\log F_D(D_1) + \log F_D(D_2) + \cdots + \log F_D(D_s)$. However, this does not take into account that there is sample variance in the *alternative hypothesis* i.e., that each $D_1, D_2, \ldots, D_s$ was drawn from a unique distribution, $F_{D_1}, F_{D_2}, \cdots, F_{D_s}$, respectively, that is estimated by sampling with replacement only from the response distributions in each partition.

And so we compute a second bootstrap, using the alternative hypothesis as the sampling frame, sampling each $D_1^*, D_2^*, \ldots, D_s^*$ directly from the bootstrapping frame associated with its partition's original sample $F_{D_1}, F_{D_2}, \cdots, F_{D_s}$ and for each sample compute its log-likelihood $\log F_D(D_1^*) + \log F_D(D_2^*) + \cdots + \log F_D(D_s^*)$ under the null hypothesis.

We then take the p-value to be the point at which the the observed test statistic is more likely under the alternative hypothesis than the null hypothesis, according to the bootstrap samples.

In each of the subfigures in of Figures 5d and 5h, the orange (leftmost) distributions are the values of the test statistic under the alternative hypothesis and the blue (rightmost) distributions are same values under the null hypothesis. The p-value is the area under the blue distribution's curve to the left of where the two curves intersect (when they intersect).

6. Experiments

Figures 5d and 5h show the results of these tests for partitioning by machine score and label class, respectively, along with the p-values associated with each test. The size of each bootstrap sample was $100K$.

	Smoothed	Collapsed	Naive	KL	Wasserstein
Precision	9/9	9/9	8/9	6/9	5/9
1	Muffin	Muffin	Muffin	Muffin	Teacher
2	Canoe	Canoe	Canoe	Canoe	Athlete
3	Chopsticks	Chopsticks	Chopsticks	Teacher	Physician
4	Chef	Chef	Chef	Graduation	Chopsticks
5	Athlete	Athlete	Athlete	Chopsticks	Coach
6	Lipstick	Coach	Coach	Chef	Funeral
7	Smile	Lipstick	Lipstick	Athlete	Smile
8	Coach	Smile	Selfie	American football	Selfie
9	Child	Child	Smile	Coach	Child
10	Selfie	Bird	Child	Lipstick	Graduation
11	Firefighter	Selfie	Firefighter	Nurse	Construction worker
12	American football	Firefighter	Bird	Selfie	American football
13	Construction worker	American football	American football	Smile	Lipstick
14	Teacher	Construction worker	Construction worker	Construction worker	Thanksgiving
15	Bird	Teacher	Funeral	Child	Firefighter
16	Funeral	Funeral	Teacher	Firefighter	Canoe
17	Physician	Physician	Physician	Physician	Bird
18	Pizza	Pizza	Pizza	Pizza	Nurse
19	Graduation	Croissant	Croissant	Funeral	Pizza
20	Croissant	Graduation	Graduation	Croissant	Muffin
21	Nurse	Nurse	Thanksgiving	Thanksgiving	Chef
22	Thanksgiving	Thanksgiving	Nurse	Bus driver	Croissant
23	Bus driver	Bus driver	Bus driver	Bird	Bus driver

Table 2: Label classes ranked by most-to-least distant from the null hypothesis distribution, according to p-value by bootstrap strategy (smoothed, naive, collapsed), KL-divergence, or Wasserstein distance. In the smoothed test, all items above line 9 reject the null hypothesis at the $p = .05$ level, with Bonferroni correction. Note that the first seven results in the first column (and more in the second and third columns) all have a p-values of less than 10^{-5}, which is beyond the precision of the bootstrap to handle. And so we used the order of the items in the KL column to settle ties in those cases.

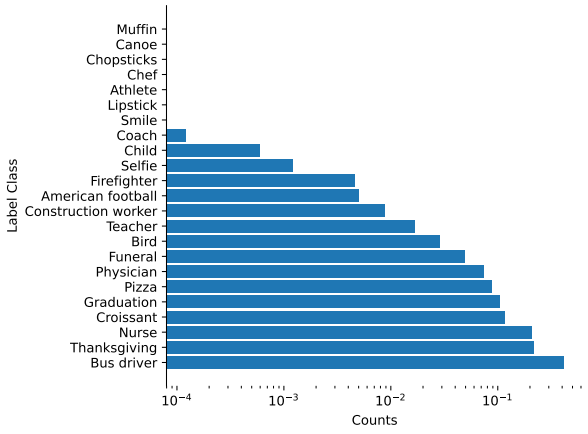


Figure 6: p-values from smoothed bootstrap samples where the test statistic is *the log-likelihood of annotator response distributions under the null hypothesis, for each class independently*, with the null hypothesis that the data was generated from a single distribution and the alternative hypothesis that data associated with each **label class** was generated from a distinct distribution. Values with no bar have estimated p-values $< 10^{-5}$. The nine classes above "Selfie" are significant at the .05 level after Bonferroni correction for 69 (3×23) tests.

As we expect, as we move from naive to collapsed to smoothed bootstrapping, the variance in each bootstrap sample increases and the null and alternative distributions move closer together. In the case of label class partitioning (Figure 5h), these trends are too small to have a measurable impact on the p-values, which were too small to measure anyway. But Figure 5d shows that for machine score the choice of bootstrap strategy makes a big difference. There, both the naive and collapsed bootstraps yield very low p-values ($p = 3 \times 10^{-5}$ and $p = .003$, respectively) and so reject the null hypothesis at very low levels, whereas the p-value for the smoothed bootstrap ($p = .3$) is too high to reject the null hypothesis at any conventional level. However, recall that we used a smoothing parameter $\alpha = 1$ that is higher than what is typically used, and smaller values can significantly decrease the p-value. For instance for $\alpha = .5$ the p-value was .18. So prior knowledge about what constitutes meaningful smoothing can be important here.

We can take these tests further and use them to discover label classes that are particularly unlikely under the null hypothesis, i.e., they are label classes that seem to invoke particularly anomalous annotator responses. Figure 6 shows the p-values for hypothesis tests using the same null and alternate hypotheses and statistical test as above, but applied one label class at a time only

to the subset of the data associated with that label.

Table 2 ranks the label classes by p-value in ascending order. For comparison to standard probability distance measures, we also show the ranked (in descending order) KL-divergence and Wasserstein distance between the class distributions (i.e., the alternative hypothesis) and the null hypothesis.

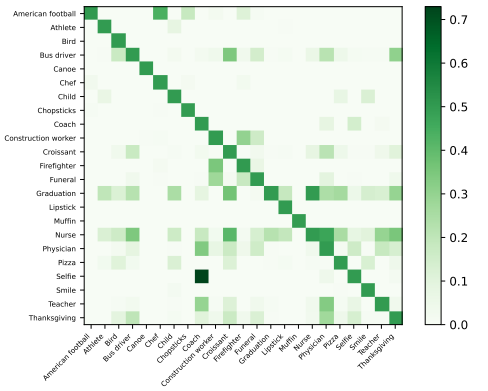
These results show that the null hypothesis distribution of all annotator distributions from all label classes combined does not represent the distributions from individual label classes. Thus it makes sense to compare label classes directly to each other. That is, we can repeat the above experiments with two label classes, where one class plays the role of the null hypothesis, the other plays the alternative hypothesis, and we use the likelihood under null hypothesis as the test statistic. In this way, p-values can be used as a similarity measure between classes. Figure 7 shows the p-values between each these pairwise tests, for smoothed bootstrapping and, for comparison purposes, KL-divergence and Wasserstein distance.

The likelihood tests above are effective for showing that conditioning on certain variables leads to meaningful distinctions in annotator response distributions. However, they tell us little about the quality of those distinctions. So, turning now on the label class condition only, we use the entropy of the annotator response distributions in each class, averaged over all of the classes.

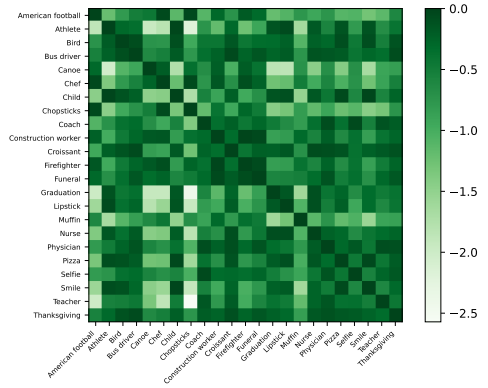
Figure 8 shows the results of these experiments. One might expect that, as one moves from naive to collapsed to smoothed that the entropy of both the null and alternate distributions would be higher and the two distributions would move closer together. But instead, somewhat unexpected things occur. First, the entropy distributions decrease slightly between the naive ($p < 10^{-5}$) and collapsed ($p < .16$) bootstraps. And then, when smoothing ($p = .0095$) is added, the entropies both increase and the distributions separate. We believe this is due to the presence in the collapsed sample of annotator distributions with no mass on certain responses (e.g., $[y_{Yes}, y_{No}, y_{\neg Know}] = [5, 0, 0]$. With no smoothing, such distributions cannot during bootstrapping generate all distributions (for instance, bootstrapping over $[5, 0, 0]$ will only ever generate $[5, 0, 0]$, whereas bootstrapping on $[1, 2, 2]$ can potentially generate any 5-annotator response). This creates biases toward these distributions, which also happen to be where most of the annotator distribution mass is located in the first stage. And so bootstrap sampling from them tends to drive entropy down. Smoothing seems to correct this, even when less smoothing is present. For instance smoothing with $\alpha = .5$ yields a p-value of $.0074$, which is still acceptably low by most standards.

7. Discussion

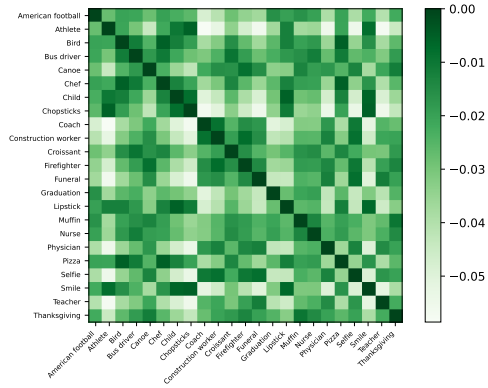
As a rule of thumb, the stronger the null hypothesis, the weaker the test. Our null hypothesis was that the an-



(a) Smoothed ($\alpha = 1.0$) bootstrap.

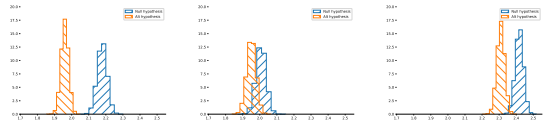


(b) -KL divergence.



(c) -Wasserstein distance.

Figure 7: Similarities between the distributions of annotator response distribution between each pair of label classes, according to p-value by smoothed bootstrapping, KL-divergence, and Wasserstein distance, respectively. For the p-value results, we zeroed out all pairs whose p-values were less than .05 after Bonferroni correction. This is because, for the purpose of hypothesis testing at the .05 level, such results are indistinguishable from those whose p-values were less than our bootstrap's precision 10^{-5}. We apply smoothing to the KL divergence results to avoid infinity results, and we take the negative of KL-divergence and Wasserstein distance.



(a) Naive bootstrap ($p < 10^{-5}$). (b) Collapsed bootstrap ($p = .16$). (c) Smoothed bootstrap ($p = .0095$).

Figure 8: Bootstrap samples where the test statistic is *the mean entropy (over all label classes) of the distribution of annotator response distributions*, with the null hypothesis that the data was generated from a single distribution and the alternative hypothesis that data associated with each *label class* was generated from a distinct distribution.

notator distributions observed in key partitions of the CATS4ML dataset were are all drawn from the same distribution. We showed that when the partitions are based on machine label alone, we may not be able to reject this hypothesis, depending on how we model variance. However, when the partitions are based on label class, differences in the annotator distributions *are* significant across multiple variants of bootstrapping.

We were able to use bootstrap-based hypothesis tests to discover annotator classes that were particularly unlikely to have been sampled from the null hypothesis, even after Bonferroni correction. We showed that the classes discovered differ slightly based on the variant of bootstrap sampling used, and differed even more from other measures of distribution similarity, including KL-divergence and Wasserstein distance.

As for how we see these methods used in the future, we found the p-values based on the log-likelihood under the null hypothesis to be useful for quantifying how different various subsamples were from each other, *in light of sampling error.* We could see it being used as an alternative to other distance or similarity measures, one that has the advantage of taking sample size into account. Often, when pairwise comparing large amounts of data, it is necessary to sparsify feature relationships, i.e., eliminate all but the most closely related pairs. Figure 7 suggests that p-values could provide a principled way to sparsify data.

This study had a number of limitations. It focused solely on differences in subsets of the dataset, which is useful for understanding the quality of data used for training and test AI systems. We would like to use similar methods to compare the performance of different AI systems on the same dataset. Such comparisons require paired hypothesis testing, which has its own complications. Hypothesis testing over items (but not annotators has long been a part of AI research (Mitchell, 1997; Dietterich, 1998; Zhang et al., 2004; Berg-Kirkpatrick et al., 2012; Søgaard et al., 2014) even if it is not as common as perhaps it should be. It is not entirely clear how much of what we learned here would apply. For instance, it would not be as easy to collapse the sampling frames in a paired setting.

We have yet to explore whether the bootstrapping methods explored here are consistent, in the sense that the expected estimates they provide approach the actual population statistics as the sample size approaches the population size. Bootstrapping, for instance fails, to have this property with respect to many statistics over long-tailed distributions.

8. Conclusion

We explore annotator responses as a sampling frame. Using the CATS4ML dataset, we show that annotator response distributions form patterns related to specific input features (*labels classes* in our case) that cannot be explained by chance, as witness by our hypothesis tests. We show that hypothesis testing can be used to identifying particularly anomalous distributional patterns and to measure the similarity between different samples in a way that accounts for sample size. We propose the log-likelihood of a sample under the null hypothesis as a used test statistic for exploration in this space. Future work will seek to extend these methods to A/B testing of AI systems that predict annotator response distributions.

Akhtar, S., Basile, V., and Patti, V. (2019). A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Basile, V. (2020). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Corani, G., Benavoli, A., Demšar, J., Mangili, F., and Zaffalon, M. (2017). Statistical comparison of classifiers through bayesian hierarchical modelling. *Machine Learning*, 106(11):1817–1837.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., and Bernstein, M. S. (2022). Jury learning: Integrating dissenting voices into machine learning models. *arXiv preprint arXiv:2202.02950*.

Klenner, M., Göhring, A., Amsler, M., Ebling, S., Tuggener, D., Hürlimann, M., and Volk, M. (2020). Harmonization sometimes harms.

Liu, T., Venkatachalam, A., Sanjay Bongale, P., and Homan, C. (2019). Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1111–1120.

Metz, C. (2019). A.I. is learning from humans. many humans. *New York Times*, August 16. `https://www.nytimes.com/2019/08/16/technology/ai-humans.html`, retrieved 5/21/2021.

Mitchell, T. M. (1997). *Machine learning, International Edition*. McGraw-Hill Series in Computer Science. McGraw-Hill.

Reidsma, D. and Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Søgaard, A., Johannsen, A., Plank, B., Hovy, D., and Alonso, H. M. (2014). What's in a p-value in NLP? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Szymański, P. and Gorman, K. (2020). Is the best better? bayesian statistical model comparison for natural language processing. *arXiv preprint arXiv:2010.03088*.

Weerasooriya, T. C., Liu, T., and Homan, C. M. (2020). Neighborhood-based pooling for population-level label distribution learning. In *ECAI 2020*, pages 490–497. IOS Press.

Welty, C., Paritosh, P., and Aroyo, L. (2019). Metrology for AI: From benchmarks to instruments. *arXiv preprint arXiv:1911.01875*.

Zhang, Y., Vogel, S., and Waibel, A. (2004). Interpreting bleu/nist scores: How much improvement do we need to have a better system? *Proc. LREC, Lisbon, Portugal, 2004*, pages 2051–2054.