Calibrate-Extrapolate: Rethinking Prevalence Estimation with Black Box Classifiers

Siqi Wu, Paul Resnick

University of Michigan, Ann Arbor
{siqiwu, presnick}@umich.edu

Abstract

In computational social science, researchers often use a pre-trained, black box classifier to estimate the frequency of each class in unlabeled datasets. A variety of prevalence estimation techniques have been developed in the literature, each yielding an unbiased estimate if certain stability assumption holds. This work introduces a framework to rethink the prevalence estimation process as calibrating the classifier outputs against ground truth labels to obtain the joint distribution of a base dataset and then extrapolating to the joint distribution of a target dataset. We call this framework "Calibrate-Extrapolate". Visualizing the joint distribution makes the stability assumption needed for a prevalence estimation technique clear and easy to understand. In the calibration phase, the techniques assume only a stable calibration curve between a calibration dataset and the full base dataset. This allows for the classifier outputs to be used for purposive sampling, thus improving the efficiency of calibration. In the extrapolation phase, some techniques assume a stable calibration curve while some assume stable class-conditional densities. We discuss the stability assumptions from a causal perspective. By specifying base and target joint distributions, we can generate simulated datasets, as a way to build intuitions about the impacts of assumption violations. This also leads to a better understanding of how the classifier predictive power affects the accuracy of prevalence estimates: the greater the predictive power, the lower the sensitivity to violations of stability assumptions in the extrapolation phase. We illustrate the framework with an application that estimates the prevalence of toxic news comments over time on Reddit, Twitter, and YouTube, using Jigsaw's Perspective API as a black box classifier.

1 Introduction

Measuring the frequency of certain labels within a data sample is a common task in many disciplines. This problem, generally called "prevalence estimation" or "quantification", has a wide range of real world applications, from quantifying the number of species in a region (Gotelli and Colwell 2001), to infected COVID-19 patients in a country (Sempos and Tian 2021), to automated accounts in a social platform (Yang et al. 2020), and to anti-social posts in an online community (Park, Seering, and Bernstein 2022). It would be ideal to manually examine every individual in the dataset.

This approach is called "census". However, census is often prohibitively expensive.

In computational social science (CSS), it is common to make use of an imperfect pre-trained classifier that assigns a label, or label probability, to each item (Yang et al. 2020; Rajadesingan, Budak, and Resnick 2021; Saha et al. 2023; Pfeffer et al. 2023). The prevalence estimation techniques then somehow account for the imperfections of the classifier. Alternative approaches are possible, such as a model trained with prevalence estimation evaluation metrics rather than individual classification accuracy as the objective function (Esuli and Sebastiani 2015). However, the classifier-aided approaches are especially attractive when the dataset is on a large scale, ground truth labels are difficult or expensive to obtain, and a high quality pre-trained classifier such as ChatGPT or Jigsaw's Perspective API (Wulczyn, Thain, and Dixon 2017) is available for use.

Extant literature on prevalence estimation with black box classifiers is plentiful and there are two useful survey articles (Forman 2008; González et al. 2017). In this paper, we do not propose any new estimation techniques. Instead, our contribution is a conceptual framework called "Calibrate-Extrapolate". It rethinks the prevalence estimation process not as merely estimating the percentage of ground truth labels but as estimating the complete joint distribution between the classifier outputs and ground truth. A calibration phase gathers ground truth labels for a small data sample, purposefully selected from a base dataset, and then estimates the joint distribution for the full base dataset. An extrapolation phase assumes that some properties of the base joint distribution are stable and infers other properties from the observed classifier output distribution of a target dataset. In both phases, the prevalence for label of interest can be read off from the inferred joint distribution of classifier outputs and ground truth labels. The Calibrate-Extrapolate framework is broadly applicable to many real world settings. It is also flexible because researchers can still customize design elements in some steps. We identify four such elements: which black box classifier to apply, which data to sample for ground truth labels, which stability assumption to make, and which prevalence estimation technique to employ.

In a real-world setting, it may be unclear which stability assumptions are reasonable to make. Thinking about prevalence estimation techniques in the Calibrate-Extrapolate

framework makes it clear what assumptions each technique requires about the stable properties (or lack thereof) between training, base, and target datasets; and how violations of those assumptions can lead to errors in the final prevalence estimate. For example, assuming only stable calibration curve between base and target datasets limits the range of possible final estimates and attenuates any observed change in the density of classifier outputs. On the other hand, assuming only class-conditional densities amplifies any observed change in the density of classifier outputs and can lead to estimates clipped at 0 or 100%. In addition, thinking about prevalence estimation in this framework clarifies the value of a more accurate classifier. While a weak classifier can still yield correct prevalence estimates in both the calibration and extrapolation phases if repeated many times, those estimates will be less certain and less robust to incorrect stability assumptions between base and target datasets.

We demonstrate the impacts of alternative design elements using simulated data, where large-scale base and target datasets are generated from known joint distributions of classifier outputs with ground truth labels. We experimented with two kinds of data generating processes (intrinsic and extrinsic), two classifiers (strong and weak), two stability assumptions (stable class-conditional densities and stable calibration curve), and two prevalence estimation techniques (distribution matching and probabilistic estimator). These yield an instructive set of comparisons that highlight both the strengths and limitations of each estimation technique.

We also apply the Calibrate-Extrapolate framework to estimating the weekly prevalence of toxic news comments on Reddit, Twitter,[1] and YouTube in year 2022. We selected Jigsaw's Perspective API[2] as our black box classifier and sent all collected comments to it for obtaining a classifier score for each. We used comments from August 2021 as the base dataset, asked crowdworkers to annotate it, and used their annotations to calibrate the Perspective API outputs. We then extrapolated the learned base joint distribution to make an inference for target datasets in later weeks. The alternative prevalence estimation techniques yield a range of estimates for each platform for each week.

In sum, the contributions of this work include:

- a framework that rethinks the prevalence estimation process as modeling the joint distribution between classifier outputs and ground truth labels in a dataset;

- experiments and arguments about how different design elements of the prevalence estimation process impact the accuracy of estimates, including guidance about avoiding clear missteps and ways to think about robustness to incorrect stability assumptions;

- an empirical prevalence estimate over time of toxic news comments on Reddit, Twitter, and YouTube.

---

[1]Twitter has recently rebranded itself as X. In this paper, we continue to use its old name, Twitter, and refer to its posts as tweets.

[2]https://perspectiveapi.com/

2 Background

For concreteness and simplicity, assume there are only two ground truth classes, which we will call positive $\oplus$ and negative $\ominus$, and positive is the class of interest. We are interested in prevalence estimation processes that make use of a black box classifier. It is black box in the sense that it was previously trained by other people on different datasets, and the prevalence estimator cannot look inside the classifier to change how it works. But we can apply the classifier to new data. Formally, given the feature vector x_i for an item i, a classifier $C(\cdot)$ would output a continuous score $C(x) \in [0, 1]$. A few widely used black box classifiers in the CSS community include Perspective API for toxicity detection (Wulczyn, Thain, and Dixon 2017), VADER for sentiment analysis (Hutto and Gilbert 2014), Botometer for online bot detection (Yang et al. 2020), and LLMs for making inferences in a variety of text classification tasks.

A naive approach of prevalence estimation would just count the number of items with classifier scores above some threshold, or sum up the classifier scores as if they were probabilities. However, these would not yield very good prevalence estimates for two reasons.

First is calibration. When a classifier outputs a score between 0 and 1, while a high value indicates higher confidence that the ground truth label is positive, it is not clear that the score is intended to be interpreted as a *calibrated* probability. It is thus not safe to assume that, given a collection of items all receiving a classifier score of 0.8, on average 80% of items would have ground truth labels to be positive. For example, prior work finds that deep neural networks tend to generate overconfident outputs. Several methods have been developed to calibrate the classifier outputs to true probabilities (Platt 1999; Guo et al. 2017).

Second is dataset shift. The classifier may have been trained on a dataset that is systematically different from the dataset for which the researchers are estimating prevalence. For example, the Perspective API was trained on comments from forums such as Wikipedia and the New York Times. However, researchers have applied Perspective API to social media comments from Reddit (Rajadesingan, Budak, and Resnick 2021), Twitter (Hua, Ristenpart, and Naaman 2020), and YouTube (Wu and Resnick 2021). The linguistic features of those comments may differ in important ways from the training data. Furthermore, even if the classifier is trained on a similar data source, it is possible that new terms might appear between the training and inference time, resulting in more frequent errors for those new terms. Dataset shift is a very important problem in machine learning literature (Moreno-Torres et al. 2012; Arora, Huang, and He 2021) and especially a concern for lexicon and rule-based classifiers such as VADER.

Prevalence estimation techniques In a recent survey of prevalence estimation (González et al. 2017), the authors categorized techniques into three groups: classify count & correct, distribution matching, and quantification learning. We will review the first two groups and rethink them in terms of the Calibrate-Extrapolate framework in Section 3. The quantification learning methods are out of scope for this pa-

per because they directly optimize a loss function for prevalence estimation evaluation metrics, which is akin to training a completely new model rather than treating the pre-trained classifier as a black box. We remark that the idea of classifier posterior calibration, which is what we do in the calibration phase of the Calibrate-Extrapolate framework, is similar to fine-tuning a pre-trained model to specific datasets and to calibrated probabilities (Guo et al. 2017).

Linking stability assumptions to data generating processes Current research in machine learning has offered a causal perspective for thinking about potential distribution shifts across datasets. This perspective considers the causal relation between a vector of covariates x and the labels y (Moreno-Torres et al. 2012; Schölkopf et al. 2012; Jin et al. 2021). Thinking of either the covariates causing labels, or the labels causing the covariates, leads to two different kinds of distribution shifts and thus two different data generating processes:

- *Intrinsic data generation → Prior probability shift*. In one conceptual model of the data generating processes, sometimes called "intrinsic data generation" (Card and Smith 2018), the label y causally determines the covariate x (causal diagram: $y \rightarrow x$). That is, we assume that y_i is drawn from some distribution and the realized y_i decides the covariates x_i. For example, when simulating, we might generate x as independent draws from different distributions for the positive or negative class. This would lead to a stable property across a base dataset B and a target dataset T: $P_B(x|y) = P_T(x|y)$. There can, however, be what is called prior distribution shift or label shift: $P_B(y) \neq P_T(y)$.

- *Extrinsic data generation → Covariate shift*. With extrinsic data generation, the covariate x causally determines the ground truth label y (causal diagram: $x \rightarrow y$). That is, we assume that x_i is drawn from some distribution and the realized x_i decides the label y_i. This would lead to a different stable property across a base dataset B and a target dataset T: $P_B(y|x) = P_T(y|x)$. There can, however, be covariate shift: $P_B(x) \neq P_T(x)$.

For this paper, we insert one more element into the causal diagrams (Figure 1): the classifier output $C(x)$. The classifier makes predictions based on the feature vectors, thus the causal direction is always $x \rightarrow C(x)$. Our framework aims to estimate the joint distributions between $C(x)$ and y, i.e., $P(C(x), y)$. We are more interested in what properties stay the same across datasets, rather than what may have shifted. To this end, we reframe the terminologies in distribution shift research to stability assumption.

- *Prior probability shift → Stable class-conditional densities*. The data generation causal diagram becomes $y \rightarrow x \rightarrow C(x)$. Using the chain rule, we derive $P(C(x)|y) = P(C(x)|x) * P(x|y)$. Because $P(C(x)|x)$ is stable (as the same classifier used across datasets) and $P(x|y)$ is stable (under intrinsic data generating process), the stability of class-conditional densities $P(C(x)|y)$ is guaranteed: $P_B(C(x)|y) = P_T(C(x)|y)$. That is, y causally decides the classifier output $C(x)$.
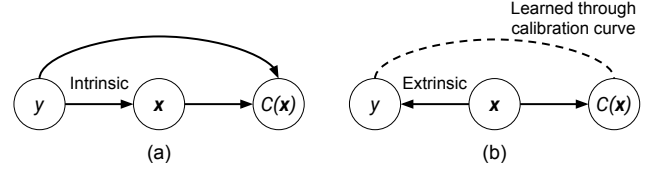

Figure 1: Causal diagrams between x, y, and $C(x)$. (a) Intrinsic data generating: y still causally decides $C(x)$. (b) Extrinsic data generating: the dashed line indicates no causal relation between y and $C(x)$. Instead, we need to learn the relation between y and $C(x)$ through fitting a calibration curve and make one additional assumption that the calibration curve function $P(y|C(x))$ is stable across datasets.

- *Covariate shift → Stable calibration curve*. The data generation causal diagram becomes $y \leftarrow x \rightarrow C(x)$. We might expect that the stability assumption generalizes, so that with this data generation process we would have $P_B(y|C(x)) = P_T(y|C(x))$. However, this is not quite true. Using the chain rule and Bayes' theorem, $P(y|C(x)) = P(y|x) * P(x|C(x)) = P(y|x) * P(C(x)|x) * \frac{P(x)}{P(C(x))}$. $P(C(x)|x)$ is still stable (i.e., the same classifier) and $P(y|x)$ is stable (under extrinsic data generating process). However, we would achieve stability between base and target datasets only if we further assume that $\frac{P(x)}{P(C(x))}$ is stable. Some prevalence estimation techniques do require that the calibration curve $P(y|C(x))$ is stable. We refer to this as the *stable calibration curve* assumption rather than refer to it as covariate shift because the covariate shift assumption is not sufficient to imply a stable calibration curve.

3 The Calibrate-Extrapolate Framework

Figure 2 illustrates the Calibrate-Extrapolate prevalence estimation framework. To present alternative prevalence estimation processes within the framework, we use a running example, estimating the fraction of Republicans among users of a U.S. social media platform at two different points in time, a base period and a target period. Assume that we have a pre-trained political leaning classifier that takes as input a corpus of text written by a person and outputs a 1 if it is sure the writer is a Republican, 0 if sure they are not, and numbers in between to express varying degrees of confidence. This yields a base dataset of people with corresponding classifier scores $C(x)$ (Figure 2A). Those classifier scores can be plotted as an empirical classifier score density curve (Figure 2B).

3.1 Calibration Phase

In the calibration phase, we select a limited sample for which we will acquire ground truth labels y based on their voter registration records (Figure 2C and D). Following the current conventional color-coding in U.S. politics, we use red for Republicans and blue for Democrats. Getting the labels may be an expensive process, requiring verifying real names and accessing records from local municipalities; hence, we
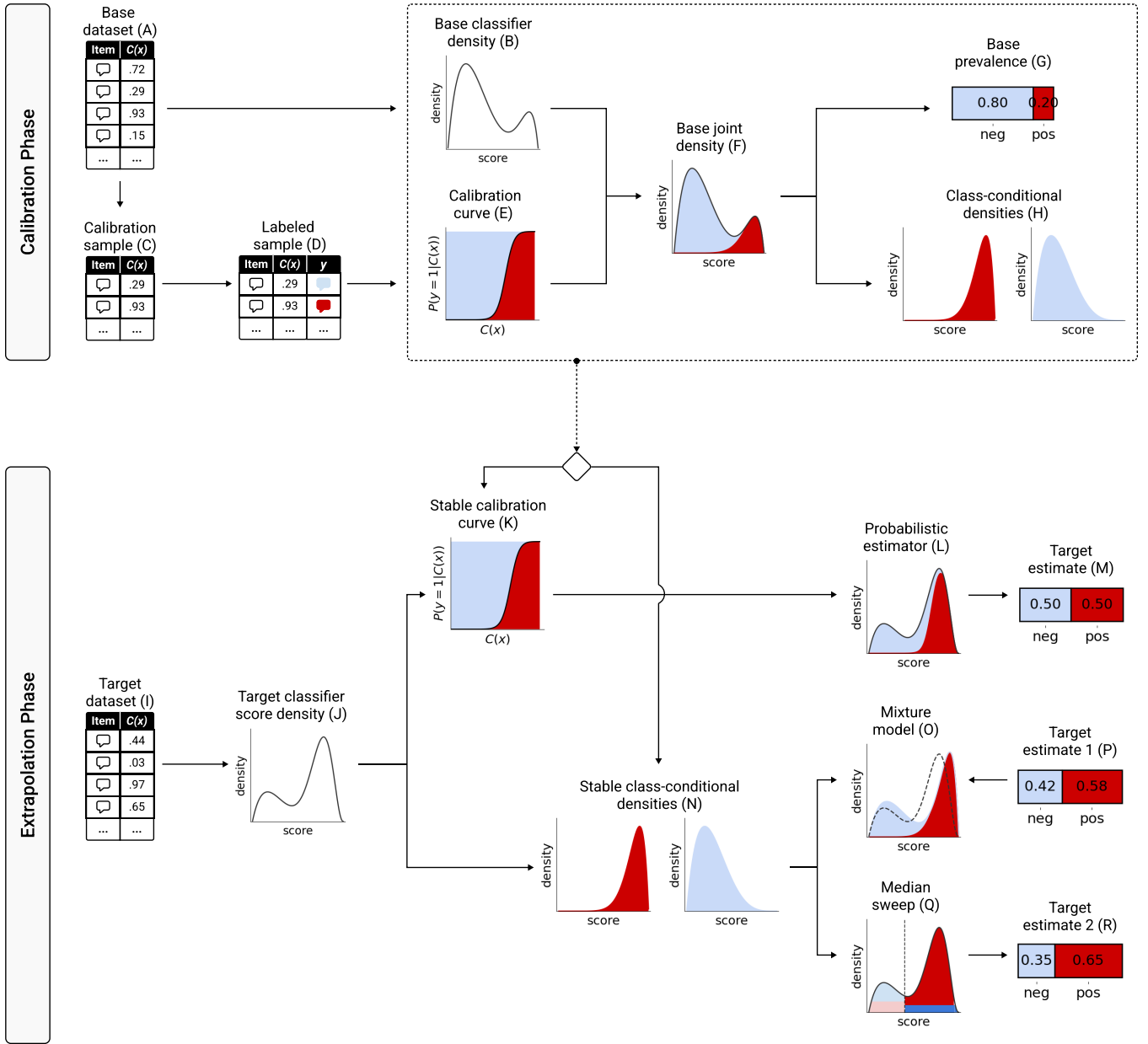
Figure 2: The Calibrate-Extrapolate framework. (top panel) The calibration phase aims to estimate the label prevalence for a base dataset B. We apply a black box classifier $C(\cdot)$ to the feature vector $\boldsymbol{x}$ of each item in B and obtain a base classifier score density (step A → B). We curate a calibration sample S from B (step A → C). We then obtain the ground truth labels for S (step C → D). Next, we fit a calibration curve function on S (step D → E). We can use the calibration curve of S and the observed classifier score density of B to compute the joint distribution between $C(\boldsymbol{x})$ and y for B (step B + E → F). This is assuming calibration curves between B and S are stable, i.e., $P(y|C(\boldsymbol{x}))_S = P(y|C(\boldsymbol{x}))_B$. From the joint distribution, we can derive the class-conditional densities for positive and negative class in B (step H) and make a prevalence estimate of B (step G). (bottom panel) The extrapolation phase aims to estimate the label prevalence in a target dataset T. We apply the same classifier $C(\cdot)$ to all items in T (step I → J). By making a stability assumption (step K, N) that links the joint distributions of B and T, we can infer the joint distribution of T (step L, O, Q) and read off the final prevalence estimate of T (step M, P, R).

carry it out only for a limited sample. The sample may be selected at random. To make the best use of a limited set of expensive ground truth labels, it can be helpful to use the classifier to select a purposive sample based on the classifier outputs. For example, it may be helpful to oversample in ranges of $C(\boldsymbol{x})$ that occur rarely, especially for highly imbalanced datasets and that imbalance reflects on the distribution of $C(\boldsymbol{x})$. Two sampling strategies can possibly optimize the efficiency of final base dataset prevalence estimate (i.e., have the smallest confidence interval around it): one is to sample equally from every $C(\boldsymbol{x})$; the other is called "Neyman allocation" (Neyman 1934).

From the labeled calibration sample, we can estimate a calibration curve, which is a function that maps the classifier outputs to calibrated probabilities (Figure 2E). There are several options for this estimation process. One is non-parametric binning: divide the $C(\boldsymbol{x})$ values into bins; compute the empirical fraction of positives (Republicans) in the bin; and treat that fraction as the calibrated probability for any $C(\boldsymbol{x})$ within the bin. Isotonic regression adds an extra constraint that calibrated probabilities must be higher for larger $C(\boldsymbol{x})$ values (Chakravarti 1989). Platt scaling makes a further parametric assumption of a smooth sigmoid curve, as we have illustrated in the figure; it is estimated through a logistic regression that treats $C(\boldsymbol{x})$ as the independent variable and the binary ground truth y as the dependent variable (Platt 1999). Finally, temperature scaling extends Platt scaling to multi-class inputs, ideal for concatenating as a calibration layer in deep neural networks (Guo et al. 2017).

Note that a classifier density function (Figure 2B) and a classifier calibration curve (Figure 2E) together define the joint distribution of classifier scores and ground truth labels. To see that, imagine using the calibration curve to color each vertical slice of the classifier density plot, yielding the more familiar representation of a joint distribution as a stacked joint density plot (Figure 2F). At low $C(\boldsymbol{x})$ values, the vertical slice is colored almost entirely blue; at high values, almost entirely red.

Alternatively, three other properties can be used to define a joint distribution: the ground truth prevalence (Figure 2G) and two class-conditional density functions, one for Republicans and one for Democrats (Figure 2H). The converse is also true. Given a different representation for the joint distribution, it is possible to read off these properties. For example, given the joint density plot (Figure 2F), the sizes of the blue and red areas under the curve determine the ground truth prevalence (G). And looking at the blue and red areas separately from the stacked joint density plot yields separate density plots for each class (H). Note that (G) gives the estimated prevalence of Republicans in the base dataset, the quantity of interest. However, other properties of the base joint distribution will be important for the extrapolation phase, it is thus helpful to think about the calibration phase as creating an estimate of the complete joint distribution.

Before turning to the extrapolation phase, it is worth reinterpreting some existing prevalence estimation techniques in terms of processes yielding the base prevalence estimate in Figure 2G. One is *calibrated probabilistic classify and count* (Bella et al. 2010; Card and Smith 2018). The calibrated part

refers to passing each item's classifier score through the calibration curve (Figure 2E) to yield a well-calibrated probability. Those probabilities are added up and the sum is divided by the total number of items, yielding the prevalence estimate. The summation of probabilities is equivalent to first counting how many items have each classifier score and then multiplying the corresponding calibrated probabilities by the frequency of classifier scores; that is what happens when we integrate across the possible density scores in (Figure 2F) to compute the size of the red area (G).

An even simpler version is an uncalibrated *probabilistic classify and count*. This technique skips the steps of sampling and labeling (Figure 2B and C) entirely and just adds up the classifier scores as if they were calibrated probabilities. This is equivalent to assuming, rather than estimating, that the calibration curve (Figure 2E) is a straight line from $(0,0)$ to $(1,1)$: $P(y = \oplus | C(\boldsymbol{x})) = C(\boldsymbol{x})$. If the classifier is indeed perfectly calibrated, the prevalence estimate will be fine; but if not, the estimate could be far off.

Other approaches binarize the classifier outputs by applying a threshold. For example, any classifier scores above 0.7 might be treated as positive and lower scores as negative. This leads to the *classify and count* and *adjusted classify and count* techniques. The unadjusted version simply treats the fraction of items with $C(\boldsymbol{x})$ scores above the threshold as the prevalence estimate. The adjusted version computes a multiplier that is applied to the estimate. In our framework, these approaches are equivalent to collapsing the calibration curve into a step function, with one universal probability for scores above the threshold and another for scores below the threshold. In the unadjusted version, one assumes perfect classification: for scores above the threshold the calibrated probability is 1 and for those below it is 0. In the adjusted version, the empirical calibration curve is computed and then it is collapsed to provide two scores; these scores correspond to the multipliers in the adjusted classify and count technique. Note that the joint distribution becomes just a confusion matrix if applying any threshold.

We also note one other point of caution when applying thresholds. It is important to first compute a continuous calibration curve as shown in Figure 2(E), and only collapse it into a binary step function by computing the red vs. blue area for classifier scores above and below the threshold. The collapsed step function treats all classifier scores as if they imply the same probability of a positive ground truth label. It is likely, however, that scores closer to the threshold are less likely to be ground truth positive than scores close to 1. This needs to be taken into account when computing the red vs. blue area for all scores above the threshold.

Another point of caution is that only calibration curve should be estimated from the calibration sample, especially when the items with ground truth labels are not selected at random from the base dataset. Purposive sampling alters the observed class-conditional densities. Imagine a scenario where items of $C(\boldsymbol{x}) \in [0.9, 1]$ are rare, and we want to include all items in that range to reduce the width of confidence intervals near it when fitting a calibration curve. If items of $C(\boldsymbol{x}) \in [0.9, 1]$ are more likely to be ground truth positive than other ranges, this will artificially increase the

density in $[0.9, 1]$ for the positive class density function.

It is also worth noting that if the researcher's goal is only to produce a single prevalence estimate, for a base dataset, a randomly selected sample would yield an unbiased estimate. Indeed, the final base prevalence estimate will be essentially the same as one would get by simply taking the fraction of positive items in the calibration sample. One way the classifier helps in making a prevalence estimate is that it can improve efficiency, allowing the same number of labels to yield a tighter confidence intervals around the base prevalence estimate. This is achieved through purposefully oversampling minority classes indicated by the classifier scores.

3.2 Extrapolation Phase

As we shall see, the real power of having a black box classifier appears when we want to reuse it to make a prevalence estimate for additional target datasets. Having estimated the joint distribution of classifier and ground truth on the base dataset in the calibration phase, in the extrapolation phase we can make some assumptions about properties of the base joint distribution that remain stable. Two lines of techniques make different stability assumptions (Figure 2K and N).

First, consider the left-side representation of a joint distribution as defined by a classifier density function (Figure 2B) and a calibration curve (Figure 2E). We assume that the classifier density function from the base joint distribution is stable and thus also applies to the target dataset. The classifier density function, however, may have changed (e.g., Figure 2J differs from B). This intuitively corresponds to an extrinsic data generating process where the features of an item x_i causally determine the ground truth label y. But as we noted in Section 2, the assumption of a stable calibration curve requires an additional assumption beyond just the intrinsic data generation.

This leads to the *probabilistic estimator* technique (Bella et al. 2010). The target classifier density function is estimated from the empirical frequency of classifier scores in a target dataset (Figure 2J). Together with the calibration curve, that fully determines the joint distribution. The joint density graph, with colorings (Figure 2L), can be derived by coloring vertical slices of target classifier score density. The prevalence estimate for the target dataset (Figure 2M) can be read off from the joint density by integration, computing the size of the blue and red areas under the curve.

Next, consider the right-side representation of a joint distribution as defined by a ground truth prevalence (Figure 2G) and the class conditional densities (Figure 2H). Here, we assume that the class conditional densities are stable. The ground truth prevalence, however, may have changed (e.g., Figure 2M, P, R differs from G). As noted in Section 2, this intuitively corresponds to an intrinsic data generation process where the ground truth label y (e.g., being a Republican) causally determines the observed features x (e.g., the text that was written), and causally determines the resulting classifier score $C(x)$ (e.g., a classifier judges it is written by a Republican).

Two techniques are based on this second stability assumption. One is called the *mixture model* (Forman 2005). It treats the observed frequency of classifier scores for the target dataset as a hint about the true label prevalence. It does a grid search over the range of possible prevalence estimates. Each possibility defines a mixture of the two class-conditional densities and thus generates an implied classifier score density function. A distance metric needs to be introduced to compare the implied classifier density to the observed classifier score density (Figure 2J). Several distance metrics are possible, one popular choice in the literature is Hellinger distance (González-Castro, Alaiz-Rodríguez, and Alegre 2013). The final target prevalence estimate is the value that minimizes the selected distance function. Note that, the optimal estimate may still yield an implied classifier score density not matching (J), illustrated as the gap between the dashed line and filled area in (Figure 2O).

The other technique that assumes stable class-conditional densities is called *median sweep* (Forman 2008). It performs a different grid search, across possible thresholds for converting continuous classifier outputs to binary positive positive or negative labels. For each possible threshold, the red class-conditional density determines the ratio of red above and below the threshold, and similarly for blue. Together with the cumulative classifier density above and below the threshold, that uniquely determines the blue vs. red fraction above and below the threshold (Figure 2Q). And that, in turns, determines the total blue and red areas, implying a target estimate (Figure 2R). There is no strong reason to prefer any one threshold over another, however, this technique resolves that by taking the median of all the estimates yielded by different choices of classifier threshold. This lacks theoretical justification, but in some empirical datasets this threshold varying approach worked well (Forman 2008; González et al. 2017).

4 Impacts of Alternative Design Elements on the Accuracy of Prevalence Estimates

This section uses simulated data to explore the impacts of different design elements in the prevalence estimation process on the accuracy of estimates. The analysis here yields insights about the impact of classifier predictive power. It also yields intuitions about the errors that will occur when there is a mismatch of data generating process and stability assumption across the base and target datasets.

Our procedure is to simulate base and target datasets from joint distributions with known parameters, run several alternative estimation processes, and compare the resulting estimates with the correct prevalence values that can be read off from the original parameters. We generate large enough datasets so that sampling variance is small; differences in the prevalence estimates are caused by alternative stability assumptions modeled into the estimation processes.

4.1 Dataset and Classifier Simulation Process

We use two data generating approaches for our simulations, because each provides scenarios that yield different insights. A joint distribution between the ground truth label y and classifier output $C(x)$ is sufficient to simulate a dataset in which every item there has a ground truth label and a classifier score. Recall from Figure 2 that a joint distribution can

be specified either by the combination of a classifier score density and a calibration curve, or by a label prevalence and the two class-conditional densities. Our first data generator assumes stable class-conditional densities and uses two different label densities for the base and target datasets. Our second data generator assumes a stable calibration curve and uses two different classifier score densities, one for base dataset and one for target dataset. Details follow.[3]

Intrinsic data generator & Stable class-conditional densities As described in Section 2, this is the intrinsic data generating process ($y \rightarrow \boldsymbol{x} \rightarrow C(\boldsymbol{x})$). We model the joint distribution as a mixture of two Beta distributions, one for positive class $\mathcal{B}_{\oplus}(\alpha_{\oplus}, \beta_{\oplus})$ and one for negative class $\mathcal{B}_{\ominus}(\alpha_{\ominus}, \beta_{\ominus})$. We specify the label probability for a base dataset $P_B(y = \oplus)$ and for a target dataset $P_T(y = \oplus)$. To simulate an item, we first make a Bernoulli draw according to the label probability, yielding a ground truth label $y_i \in \{\oplus, \ominus\}$. We then generate a classifier score $C(\boldsymbol{x}) \in [0, 1]$ as a draw from the corresponding Beta distribution.

Note that the four parameters that define the two Beta distributions are kept fixed between the base and target datasets. The only parameter that changes between B and T is the label probability. Thus, with this approach we can generate a dataset from the following five parameters: ($\underbrace{\alpha_{\oplus}, \beta_{\oplus}, \alpha_{\ominus}, \beta_{\ominus}}_{\text{class-conditional densities}}$, $\underbrace{P(y = \oplus)}_{\text{label probability}}$). The invariant property is bolded, while the varying property is italicized.

In our simulations using this approach, we always use 20% positive for the base dataset, $P_B(y = \oplus) = 0.2$, and 60% positive for the target dataset, $P_T(y = \oplus) = 0.6$. For the two Beta distributions, we design two scenarios:

- strong classifier: $\mathcal{B}_{\oplus}(10, 2), \mathcal{B}_{\ominus}(2, 5)$
- weak classifier: $\mathcal{B}_{\oplus}(7, 6), \mathcal{B}_{\ominus}(2, 5)$

These two intrinsic data generators are illustrated in the first plots in rows 1 and 2 of both Table 1 and Table 2. Note that we have superimposed the two class-conditional density functions. This is clear in row 1, where there is little overlap, indicating that the classifier has high predictive power; while the substantial overlap in row 2 shows a weak classifier that does not discriminate well between positives and negatives.

Extrinsic data generator & Stable calibration curve As described in the Section 2, this corresponds to the extrinsic data generating process ($y \leftarrow \boldsymbol{x} \rightarrow C(\boldsymbol{x})$), but requires a somewhat stronger assumption, that the calibration curve $P(y|C(\boldsymbol{x}))$ is the same for the target dataset as for the base dataset. In this approach, we simulate an item as a draw from a classifier score density function, which we model as a mixture of two Beta distributions with five parameters, $\lambda \mathcal{B}_1(\alpha_1, \beta_1) + (1 - \lambda)\mathcal{B}_2(\alpha_2, \beta_2)$. This yields a realized classifier score, $C(\boldsymbol{x}_i) \in [0, 1]$. To generate a ground truth label y_i, we first find the calibrated probability associated with that classifier score, using a pre-defined calibration curve $Calib(\cdot)$. We then draw the ground truth label y_i as a Bernoulli draw with the computed calibrated probability

$Calib(C(\boldsymbol{x}_i))$. Our calibration curve is specified by two parameters, w and b, as would result from a logistic regression, i.e., formula of Platt scaling (Platt 1999):

$$Calib(w, b) = \frac{1}{1 + e^{w \cdot C(\boldsymbol{x}) + b}}$$

We use the same calibration curve, with fixed parameters w and b, for both the base and target datasets, but vary the classifier score density curve. Thus, we have a total of seven parameters to generate a dataset in an extrinsic fashion, ($\underbrace{w, b}_{\text{calibration curve}}$, $\underbrace{\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda}_{\text{classifier score density}}$). The invariant property is bolded, while the varying property is italicized.

In our simulations using this approach, we design the following classifier density score functions:

- base dataset: $C(\boldsymbol{x}) \sim 0.2\mathcal{B}_1(10, 2) + 0.8\mathcal{B}_2(2, 5)$
- target dataset: $C(\boldsymbol{x}) \sim 0.6\mathcal{B}_1(10, 2) + 0.4\mathcal{B}_2(2, 5)$

These base and target classifier score densities are the second and third plots in rows 3 and 4 of Table 2. We simulate a strong and a weak classifier, and keep their parameters fixed between the base and target datasets:

- strong classifier: $Calib(w = 25, b = -15)$
- weak classifier: $Calib(w = 0.5, b = -1)$

These two intrinsic data generators are illustrated in the first plots in rows 3 and 4 of both Table 1 and Table 2. The strong classifier has a steep slope with a calibrated probability near 0 or 1 for most classifier scores. On the other hand, the weak classifier has similar calibrated probabilities for low and high scores, suggesting little discriminating power.

For each combination of (intrinsic or extrinsic) data generator and (strong or weak) classifier, we generate a 20,000-item base dataset and a 20,000-item target dataset. Then, we apply the Calibrate-Extrapolate framework as described in Figure 2. For the calibration phase, we curate a calibration sample of up to 2,000 items by sampling up to 200 from each of ten classifier scores strata (0-10%; 10-20%; etc.), less if there are fewer than 200 items in the respective bin. We use Platt scaling method to estimate the calibration curve. For the extrapolation phase, we make prevalence estimates using two alternative techniques, probability estimator and mixture model. To construct confidence intervals, we repeat the entire process of calibration and extrapolation for 1,000 times, taking different calibration samples each time. We report the mean of the bootstrapped estimates and treat the middle 95% as the confidence intervals, following the percentile bootstrap procedure (Diciccio and Romano 1988).

4.2 Impacts of Classifier Predictive Power in the Calibration Phase

Table 1 shows the experiment results for the calibration phase for four groups under different data generating processes and different classifier predictive powers. We find that, from calibration sample to base dataset, one has to assume a stable calibration curve because the sampling is conducted based on $C(\boldsymbol{x})$. Taking the example of Neyman allocation, the sampling within each stratum is still random; but

---

[3]Our data simulation and prevalence estimation tool is publicly available at https://github.com/avalanchesiqi/pyquantifier

Table 1:

Id	Generating properties of base joint distribution	Estimated properties of base joint distribution	Prevalence estimator calib. probabilistic est.	no calib.	True
1-intrinsic-strong			**20.40%** **(19.86%, 20.91%)**	39.68%	20.00%
2-intrinsic-weak			**19.71%** **(18.28%, 21.21%)**	35.36%	20.00%
3-extrinsic-strong			**24.14%** **(23.55%, 24.67%)**	39.41%	23.13%
4-extrinsic-weak			**31.25%** **(28.89%, 33.17%)**	39.45%	31.22%

Table 1: Calibration phase experiment results. For experiment group 1-4, we specify two properties (first column plots) that are sufficient to define a joint distribution between $C(\boldsymbol{x})$ and y. We can infer other properties of the same base joint distribution and visualize them in the second column plots. Numbers in the bracket indicate 95% confidence intervals via bootstrapping.

Id	Stable generating properties	Non-stable generating properties $(B \to T)$	Inferred properties $(B \to T)$	Prevalence estimator mixture model	calib. probabilistic est.	True
1-intrinsic-strong				**60.21%** **(60.00%, 61.00%)**	<u>52.63%</u> (51.60%, 53.30%)	60.00%
2-intrinsic-weak				**60.51%** **(57.00%, 63.00%)**	23.91% (22.57%, 25.72%)	60.00%
3-extrinsic-strong				<u>63.11%</u> (62.48%, 64.00%)	**58.16%** **(57.61%, 58.69%)**	59.72%
4-extrinsic-weak				97.97% (91.75%, 100.00%)	**33.32%** **(31.19%, 35.51%)**	33.38%

Table 2: Extrapolation phase experiment results. We keep one property fixed between the base and target datasets (first column plots). We vary the non-stable generating property (second column plots). We could make an alternative stability assumption and we show the inferred property needed for the corresponding estimation technique (third column plots). Numbers in the bracket: 95% confidence intervals. Bolded numbers: estimates under the correct stability assumption.

over the entire span of $C(\boldsymbol{x})$, it is not. The class-conditional densities would thus change from base dataset to calibration sample, invaliding prevalence estimation techniques relying on such assumption. Passing the raw classifier scores to a calibration curve learned from the annotated sample, and then summing up the calibrated probabilities can accurately recover the ground truth prevalence, regardless of the data generator or classifier power (**bolded numbers**); while simply summing up the uncalibrated classifier scores yield bad estimation performance. Nevertheless, weak classifiers still have wider confidence intervals in the repeated measures.

4.3 Sensitivity to Violations of Stability Assumptions in the Extrapolation Phase

Table 2 shows the experiment results for the extrapolation phase for four pairs of (base $\to$ target) comparisons. We find

that, from base dataset to target dataset, one has to choose the stability assumption compatible with the data generating process to make a good prevalence estimate. That is, mixture model for intrinsic data generator and calibrated probabilistic estimator for extrinsic data generator. If the stable property between base and target dataset is correctly picked, the estimates will be accurate, regardless of classifier power (**bolded numbers**). Second, even under a wrong stability assumption, a strong classifier can still make decent estimates, as shown by the underlined numbers. This clarifies the value of training a high quality classifier, for its robustness to violations of assumptions in prevalence estimation.

4.4 Discussion

We remark a crucial difference between the sampling approach in our framework and in previous quantification lit-

erature (Forman 2005; González et al. 2017). In those works, the authors started with the ground truth labels y, and artificially created test datasets of different levels of prevalence conditioned on y. For example, quoting the experiment protocol from (Forman 2005), *"we randomly drew 200 positives and 1000 negatives from each benchmark classification task as the maximum training set... we randomly removed positives or negatives to achieve various desired testing class distributions"*.

Sampling based on the ground truth is attractive, because it would preserve the class-conditional density function for each class, thus enabling estimation techniques that rely on the stable class-conditional density assumption. However, sampling based on ground truth y is less realistic in social media research because it is difficult to obtain the ground truth for every item. Instead, CSS researchers are usually more interested in making unbiased prevalence estimates for large-scale social data and monitoring the trend of prevalence. In addition, they usually have access to high quality pre-trained models. They can use these models to strategically select a better sample for ground truth labels, for instance, oversampling rare class to overcome the data imbalance issue. This can improve the performances in both classification and prevalence estimation models built from the calibration sample.

In addition, assuming only stable calibration curve between the base and target datasets limits the range of possible final estimates. For example, if the learned calibration curve for base dataset $P_B(y|C(\boldsymbol{x}))$ is bounded between 20% and 60%, it means that the estimated prevalence for any target datasets would never exceed 60%, nor fall short of 20%. On the other hand, assuming only class-conditional densities amplifies can lead to estimates clipped at 0 or 100%. For example, if adding any fractions of negative class density to the positive class density makes the mixed classifier score density more dissimilar to the observed classifier score density in a target dataset, the estimator would output 100%.

5 Application: How Many Toxic Comments Are Posted on Social Media Everyday?

We apply the Calibrate-Extrapolate framework to a real world application that estimates the extent to which online comments on each week's news stories would be perceived as toxic by the general public. Separate estimates are made for three platforms: Reddit, Twitter (X), and YouTube. For each, we collected human labels for one calibration sample, then conducted an extrapolation phase for each week for one year. We performed two estimations assuming either stable calibration curves (calibrated probabilistic estimator) or stable class-conditional densities (mixture model), in order to illustrate sensitivity to different stability assumptions.

5.1 Public News Comments Dataset

Details of the data collection are described in a separate report.[4] We report a summary here. The contribution of this paper is to explore techniques for estimating prevalence in datasets like this one; the dataset itself is not a contribution.

Each day, the 1,000 most engaging URLs on Facebook and on Twitter were collected, as reported by NewsWhip.[5] Following the semi-supervised approach used in (Bakshy, Messing, and Adamic 2015), a machine learning classifier was trained to identify the URLs that reference "hard news" (e.g., politics, economics) as opposed to softer news (e.g., sports, music). For each hard news URL, a search was conducted for posts on Reddit and Twitter that contain the URL and YouTube videos whose descriptions match the news story's text. All comments on the matched posts and videos were scraped, excluding deleted, moderated, and bot comments on Reddit; retweets and quoted tweets on Twitter.

To make the final metrics comparable across platforms, we selected an equal number of comments from each platform about each hard news URL, the minimum number available across the three platforms. This sampling strategy avoids the risk that controversial stories, whose comments are frequently toxic, might garner more comments on one platform than on another. We did this to ensure that differences in the measured frequency of toxic comments reflect different modes of expression about the same stories rather than differences of popularity across platforms.

We report here on target datasets of comments from 2022. There were 15,564 hard news URLs having at least ten comments on all three platforms, about 42.6 per day. On average the collection includes 5,631 distinct comments per day on Reddit, 5,355 on Twitter, and 4,271 on YouTube.

The comments posted during August, 2021 form our base dataset. A purposive sampling process was conducted, loosely based on Neyman allocation, yielding 1,144 Reddit comments, 1,154 Twitter replies, 1,162 YouTube comments. These were annotated by Amazon Mechanical Turk (MTurk) workers, who were asked to label whether a comment was toxic, with the same definition used when collecting training data for the Perspective API (Wulczyn, Thain, and Dixon 2017)—"a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion". Five workers labeled each comment; the ground truth label in this dataset is the majority vote.

We limited access to the U.S. residents who had completed at least 1,000 Human Intelligence Tasks (HITs) with $\geq 98\%$ acceptance rate. They first had to complete a qualification task, in which we provided the concept definitions, annotation instructions, five practice comments with labeling answers given by domain experts, and four questions with clear answers. To qualify, a worker had to answer all four questions correctly. Qualified workers were then permitted to label up to 100 comments, in 10-comment batches. HITs were priced with a target hourly pay rate of $15, which is somewhat higher than the typical labor rate on MTurk.

5.2 Prevalence Estimates Using the Perspective API

For the black-box classifier, we used the Perspective API. We detected that the Perspective API changed significantly in May 2022. For many comments, the Perspective API returned a lower score after that date. Thus, we can think there
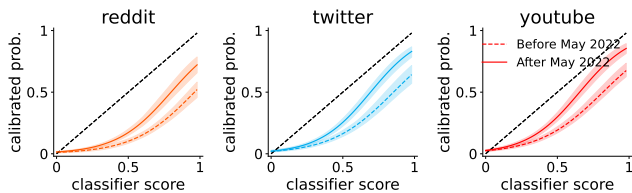
---

[4]https://csmr.umich.edu/projects/hot-speech/

[5]https://www.newswhip.com/

Figure 3: Calibration curves for two versions of Perspective API, both against the calibration sample from August 2021.
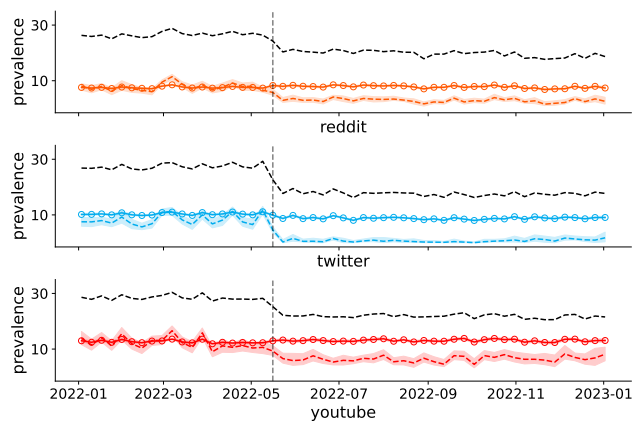

Figure 4: Toxicity prevalence estimates using three alternative estimation techniques. Colored circled: assumed stable calibration curve; colored dashed: assumed stable class-conditional densities; black dashed: without calibration. The vertical line indicates the date when the Perspective API switched to a classifier that output generally lower scores.

being two different classifiers, both called Perspective API.

Figure 3 shows the calibration curves for the two classifiers, estimated from the labeled calibration sample, using the Platt scaling technique. Since scores were generally lower with the later classifier, the calibrated probability associated with any given score was higher after May 2022. Note that in both periods, however, the calibration curves lie entirely below the 45 degree line. This indicates that Perspective API scores are systematically higher than the true calibrated probability: for example, on comments where the Perspective API outputs 0.5, less than half of comments were labeled as toxic by MTurk workers.

Also note that the calibration curves, while similar, are not exactly the same for the three platforms. The calibrated probability of being labeled toxic was higher for a comment from YouTube receiving a Perspective API score of 0.5 than for a comment from the other platforms with the same score. This underscores the need to calibrate a classifier separately when using it on datasets that have distinct error profiles.

Figure 4 shows our estimates of perceived toxic comments for each week in 2022, for each of the three platforms. The black dashed line uses the Perspective API scores as if they were calibrated probabilities, effectively following the unadjusted "probabilistic classify and count" approach. The vertical dashed line shows when the classifier changed, yielding lower prevalence estimates. However, since classifier scores still were higher than their calibrated values, the estimates without calibration are consistently too high for all three platforms, by very large margins.

The two colored lines for each platform show the results of extrapolation phase estimation assuming either a stable calibration curve from base to target dataset (calibrated probabilistic estimator technique) or stable class-conditional densities (mixture model approach). Given the sizes of calibration sample and each week's datasets, the confidence intervals due to sampling error are fairly tight. However, the difference in estimates using the two techniques, based on different stability assumptions, yields much larger differences. Assuming stable class-conditional densities leads to more extreme estimates: when the target period scores are lower than the base period, the prevalence estimate moves lower by even more. This can even lead to different conclusions about the ordering of platforms. For example, assuming stable class-conditional densities leads to a conclusion that Twitter had lower toxicity than Reddit towards the end of 2022, but assuming a stable calibration curve leads to a conclusion that Twitter had higher toxicity than Reddit.

Note that the calibrated probabilistic estimator technique seems to have handled the May changeover in the Perspective API better than the other two approaches. While the other approaches show sharply lower prevalence estimates as soon as the Perspective API gave lower scores, with the calibrated probabilistic estimator, the estimated prevalence was slightly higher on Reddit and YouTube in the weeks after the change but slightly lower on Twitter. Since it is unlikely that there was a major and long-lived change in the actual toxicity on the platforms immediately following the Perspective API change, the fact that the probabilistic estimator technique gives similar results using either version of Perspective API is one reason to think that its assumptions are better suited to this toxicity prevalence estimation task.

6 Conclusion

We have presented the "Calibrate-Extrapolate" framework. It rethinks the prevalence estimation process as calibrating the classifier outputs against ground truth labels to obtain the joint distribution of a base dataset and then extrapolating to that of a target dataset. Visualizing the joint distributions makes the stability assumption needed for a prevalence estimation technique clear and easy to understand.

In the calibration phase, all techniques require only a stable calibration curve between a labeled calibration sample and the full base dataset. Purposive sampling based on the classifier outputs does not lead to a violation of that requirement and can improve the efficiency of prevalence estimation for a base dataset, providing tighter confidence intervals for estimates based on a limited number of labeled items. Beyond that efficiency improvement, however, a black-box classifier provides limited additional value in making a prevalence estimate for a single base dataset.

The real value of the black box classifier emerges in the extrapolation phase, where the classifier is applied to a target

dataset but no further ground truth labels are collected. All prevalence estimation techniques for an extrapolation phase with no new labels depend on borrowing some property from the base dataset joint distribution, effectively assuming stability of that property from the base to the target dataset. Some techniques assume a stable calibration curve while some assume stable class-conditional densities.

This framework provides a recipe for generating simulated datasets by specifying base and target joint distributions and using them as generators. By making instructive choices for these distributions, in Section 4 we illustrated how estimation techniques can yield incorrect estimates when their stability assumptions are violated.

Section 5 applied two of the techniques, as well as a benchmark that eschews calibration entirely, to a real world collection of 52 target datasets, one for each week in a year. There, the uncertainty due to sampling error was dwarfed by the uncertainty due to the choice of stability assumptions inherent in the techniques.

Limitations and future directions   The biggest limitation of the framework is that it does not provide much guidance about how to decide *which* stability assumption is appropriate for particular real world scenarios. We provided some guidance, by connecting the stability assumptions to different data generating models in Section 2. Some real world scenarios are more naturally described as intrinsic or extrinsic data generating processes. For example, some medical settings seem to have intrinsic data generating (a disease causes symptoms). Many social media settings, on the other hand, seem to have extrinsic data generation (the words writers choose cause readers to see messages as toxic or not). However, we point out that even if the same words are always perceived the same way, and classified the same way, that is not enough to ensure that the calibration curve for a classifier will be stable between base and target datasets.

In reality, there may be both prior probability shift and covariate shift between base and target datasets. And there can also be concept drift, where the ground truth labels for items change between datasets. A promising area for future research, then, would be to develop guidance on how to diagnose when it is appropriate to take a classifier calibration curve estimated from one dataset and apply it to another. One approach would be to collect a second calibration sample from one of the target datasets, estimate another joint distribution, and see which properties actually were stable between the base joint distribution and that target joint distribution. Properties that were stable between those two might be reasonable to assume also stable for other target datasets that are constructed in a similar way.

Despite this limitation, the Calibrate-Extrapolate framework provides conceptual clarity about the logic behind alternative prevalence techniques. Researchers and practitioners who think in terms of this framework will ask themselves the right questions about the techniques they would apply and can avoid many potential pitfalls that await the unwary prevalence estimator.

References

Arora, U.; Huang, W.; and He, H. 2021. Types of Out-of-Distribution Texts and How to Detect Them. In *EMNLP*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*.

Bella, A.; Ferri, C.; Hernández-Orallo, J.; and Ramirez-Quintana, M. J. 2010. Quantification via Probability Estimators. In *ICDM*.

Card, D.; and Smith, N. A. 2018. The Importance of Calibration for Estimating Proportions From Annotations. In *NAACL*.

Chakravarti, N. 1989. Isotonic Median Regression: A Linear Programming Approach. *Mathematics of Operations Research*.

Diciccio, T. J.; and Romano, J. P. 1988. A Review of Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Esuli, A.; and Sebastiani, F. 2015. Optimizing Text Quantifiers for Multivariate Loss Functions. *ACM Transactions on Knowledge Discovery from Data*.

Forman, G. 2005. Counting Positives Accurately Despite Inaccurate Classification. In *ECML*.

Forman, G. 2008. Quantifying Counts and Costs via Classification. *Data Mining and Knowledge Discovery*.

González, P.; Castaño, A.; Chawla, N. V.; and Coz, J. J. D. 2017. A Review on Quantification Learning. *ACM Computing Surveys*.

González-Castro, V.; Alaiz-Rodríguez, R.; and Alegre, E. 2013. Class Distribution Estimation Based on the Hellinger Distance. *Information Sciences*.

Gotelli, N. J.; and Colwell, R. K. 2001. Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness. *Ecology letters*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *ICML*.

Hua, Y.; Ristenpart, T.; and Naaman, M. 2020. Towards Measuring Adversarial Twitter Interactions Against Candidates in the US Midterm Elections. In *ICWSM*.

Hutto, C.; and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *ICWSM*.

Jin, Z.; von Kügelgen, J.; Ni, J.; Vaidhya, T.; Kaushal, A.; Sachan, M.; and Schoelkopf, B. 2021. Causal Direction of Data Collection Matters: Implications of Causal and Anti-causal Learning for NLP. In *EMNLP*.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; and Herrera, F. 2012. A Unifying View on Dataset Shift in Classification. *Pattern recognition*.

Neyman, J. 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*.

Park, J. S.; Seering, J.; and Bernstein, M. S. 2022. Measuring the Prevalence of Anti-Social Behavior in Online Communities. *ACM on Human-Computer Interaction*, (CSCW).

Pfeffer, J.; Matter, D.; Jaidka, K.; Varol, O.; Mashhadi, A.; Lasser, J.; Assenmacher, D.; Wu, S.; Yang, D.; Brantner, C.; et al. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. In *ICWSM*.

Platt, J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in large margin classifiers*.

Rajadesingan, A.; Budak, C.; and Resnick, P. 2021. Political Discussion Is Abundant in Non-political Subreddits (And Less Toxic). In *ICWSM*.

Saha, P.; Garimella, K.; Kalyan, N. K.; Pandey, S. K.; Meher, P. M.; Mathew, B.; and Mukherjee, A. 2023. On the Rise of Fear Speech in Online Social Media. *PNAS*.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On Causal and Anticausal Learning. In *ICML*.

Sempos, C. T.; and Tian, L. 2021. Adjusting Coronavirus Prevalence Estimates for Laboratory Test Kit Error. *American Journal of Epidemiology*.

Wu, S.; and Resnick, P. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *ICWSM*.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *TheWebConf*.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *AAAI*.