The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

Learning to Predict Population-Level Label Distributions

Tong Liu,[1] **Akash Venkatachalam,**[1] **Pratik Sanjay Bongale,**[2] **Christopher M. Homan**[1]
[1]Rochester Institute of Technology
Rochester, New York
tl8313|av2833@rit.edu, cmh@cs.rit.edu
[2]Gleason Corporation
Rochester, New York
pbongale@gleason.com

Abstract

As machine learning (ML) plays an ever increasing role in commerce, government, and daily life, reports of bias in ML systems against groups traditionally underrepresented in computing technologies have also increased. The problem appears to be extensive, yet it remains challenging even to fully assess the scope, let alone fix it. A fundamental reason is that ML systems are typically trained to predict one correct answer or set of answers; disagreements between the annotators who provide the training labels are resolved by either discarding minority opinions (which may correspond to demographic minorities or not) or presenting all opinions flatly, with no attempt to quantify how different answers might be distributed in society. Label distribution learning associates for each data item a probability distribution over the labels for that item. While such distributions may be representative of minority beliefs or not, they at least preserve diversities of opinion that conventional learning hides or ignores and represent a fundamental first step toward ML systems that can model diversity. We introduce a strategy for learning label distributions with only five-to-ten labels per item—a range that is typical of supervised learning datasets—by aggregating human-annotated labels over multiple, similarly rated data items. Our results suggest that specific label aggregation methods can help provide reliable, representative predictions at the population level.

1 Introduction

The goal of many supervised learning problems is to map each given data item to a single (or set of, but in any case, deterministic) label(s) according to some standard of ground truth. However, many real-world problems—such as those related to color, pain, taste, level of danger, or qualitative analysis—have different answers depending on whom is asked, even when the domain of answers is fixed (i.e., *closed domain*) or more than one answer is allowed (i.e., *multilabel*). In such cases, a single (set of) label(s) does not meaningfully solve the problem, or may hide important dissenting beliefs or opinions. Yet the impact of AI agents that fail to recognize diversity in a representative fashion ranges from banal to harmful on a societal level.

Label distribution learning (LDL) is a recent approach that replaces the goal of predicting, for each data item, a single (set of) label(s) with the more challenging and complex task of predicting a probability distribution (known as a *label distribution*) over the label choices (Geng 2016). A growing body of work has used this approach, e.g., to predict beauty in images (Ren and Geng 2017) and rate movies (Geng and Hou 2015). Until now, prior work has focused broadly on the problems that distinguish LDL from other forms of probabilistic learning. We focus here on population-based LDL (PLDL), the special case of when the goal is to predict the distribution of beliefs in a population of human annotators about the best label to associate with each data item.

A major resource bottleneck in PLDL is the quantity of human annotations needed. For any large population of labelers, any lone data item x, and any question posed of x to the labelers, the number m of labels needed to estimate (i.e., taken as a sample of) the underlying population's true distribution of beliefs about x is rather large, depending on the size of the label space and desired confidence/significance level. Meanwhile, the number of data items n needed for supervised learning normally runs into the thousands. Thus, taken independently, the total number $m \times n$ of human labels needed for training on label distributions grows quadratically, and can easily run into the millions.

Our main contribution is a new algorithmic framework for reducing the total number of human labels needed per data item, by pooling together the labels of data items *determined by clustering* ***in the space of label distributions*** to be similarly rated. Figure 1 illustrates the main idea behind this algorithm.

Specifically, we:

1. Establish the premise for our proposed approach through a real-world example where there is substantial disagreement over the annotators' interpretations of 50 data items in a common social domain, but where the label distributions appear visually in histogram to cluster into a limited number of distinct classes.
2. Introduce an algorithmic framework for label distribution learning on as few as five-to-ten labels per data item that involves an unsupervised learning phase to yield hidden

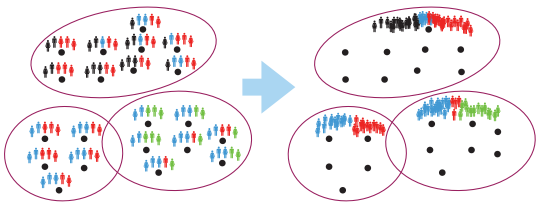

Figure 1: The main algorithmic idea this paper explores. The black dots represent data items. (Left:) Five labelers annotate each data item, where the color of the person indicates the label that person chose. If we view these five labels as a sample of the underlying population's beliefs, the sample size is probably too small for there to be much confidence in the sample. (Right:) We cluster together (indicated by the circles) similar rater response items, and then pool together all the labels in each cluster into a single, larger sample which, according to our strategy, is a good representation of—and thus label distribution for—the population-level beliefs about each item in the cluster.

classes of similarly-rated data items and assigns to each class an aggregated label distribution, followed by a supervised learning phase based on the labels the unsupervised phase produces.

3. Show that, for larger label spaces, predictions based on unsupervised learning models that use our clustering strategy outperform those that do not, thus providing supervised learning validation for our approach.
4. Perform our analysis on natural language data. This is the first explorations of LDL on linguistic data from social media (Shirani et al. (Shirani et al. 2019) also use LDL for language processing).

2 Related Work

Disagreement in human labeling tasks for supervised learning is widely studied as a common problem in its own right (e.g., (Dawid and Skene 1979)). Snow et al. (2008), in a study on using multiple crowdsourced annotators to approximate the performance of experts, noted that individuals (including experts) tend to have personal biases, but that multiple annotators may contribute to diversity, thus reducing individual annotator bias (see also (Callison-Burch 2009)). However, there is still an underlying assumption that a correct answer exists, even if it can never be directly confirmed.

Recent work has recognized the value of preserving subjectivity and ambiguity in data collection from human annotators. Aroyo and Welty show in a semantic parsing task that crowdworkers, when they agree with each other, can perform at a level comparable to domain experts, and when they disagree it is often for good reason, and in fact usually more desirable than collapsing to a single label (Aroyo and Welty 2014). Schaekermann et al. (2016) describe a framework for identifying unresolvable annotator disagreement.

Chen et al. (2018) argue persuasively that to a wide spectrum of social scientists the volume of unstructured data available for qualitative analysis generated by social media is so great that automated methods like machine learning are needed to keep up. They also argue that preserving annotator disagreement is essential to applying qualitative methodologies like grounded theory at scale.

Learning over probability distributions has a long history (e.g., (Sheng, Provost, and Ipeirotis 2008)). While **label distribution learning** (LDL) adopts many of the same algorithmic approaches from this body of work it differs from conventional learning (a) in conventional probabilistic learning probability is used to model uncertainty; in LDL probabilities model ground truth. Thus (b) while conventional probabilistic learning evaluates performance in terms of accuracy, precision, and recall (even though probabilistic measures may be used as loss functions during training) etc., in LDL performance is measured in terms of functions, such as Kullback-Leibler (KL) divergence, that operate directly on probabilities.

Geng pioneered the systematic study of label distribution learning (Geng 2016), where the objects to be predicted are probability distributions over labels/classes. He and colleagues studied applications of LDL in many settings, some of which are related to predicting population-level distributions (Geng and Hou 2015; Geng, Wang, and Xia 2014), while others are not (Gao et al. 2017). Nearly contemporary work to ours has extended the maximum entropy models in (Geng 2016) to account for covariance in the label distribution space (Jia et al. 2018).

Several of these studies acknowledge the difficulty of obtaining valid label distributions that represent the underlying beliefs of human annotators; in fact, most of them were based on data and labels originally collected for the purpose of conventional (i.e., non-probabilistic labels) supervised learning problems. However, this line of research has thus far assumed that the label distributions obtained are equal to ground truth, i.e., without questioning the statistical validity of the data, even though the sample size of the labels for each item is small.

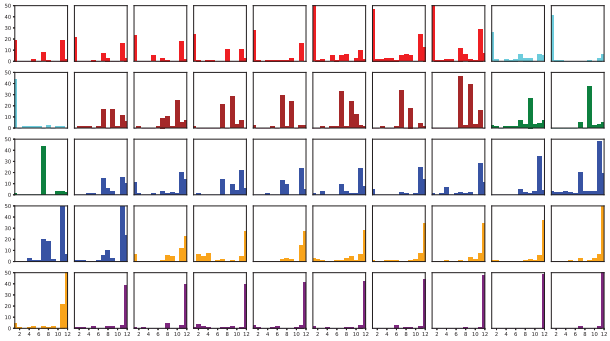

Figure 2: Each histogram above represents the label distribution of a lone data item in the jobQ3MT+ data set (Section 4.1). The X-axis ranges from 1 to 12, matching the Q3 choices in Figure 3. The Y-axis denotes the label counts. Similar distributions are grouped by color: 1-8 red, 9-11 cyan, 12-18 brown, 19-21 green, 22-32 blue, 33-41 orange, and 42-50 purple.

.

A number of research areas are related to LDL. In **multilabel learning** (Zhang and Zhou 2014), each data item is associated with multiple labels. However, it does not typically distinguish between multiplicity due to disagreement (where different annotators might believe that only one label is correct, but disagree on which one), ambiguity (where an annotator might believe multiple labels are valid), or uncertainty. Such distinctions may have significant social impacts, especially when disagreements fall along crucial demographic boundaries or indicate important but opposing perspectives that should just be preserved in the machine learning predictive models. Moreover, there are settings where label distributions are important but multilabel approaches do not naturally apply, such as when the prediction is ordinal (e.g., Likert-scaled) or real-valued. We are interested in capturing the diversity of beliefs across a population, where each member of the population may only associate a data item with one (set of) label(s), but different people may disagree on which ones. **Label propagation** is a class of semi-supervised learning algorithms that propagate labels from labeled to unlabeled data points and exploit correlations and interactions between data items in terms of feature neighborhoods (Zhu and Ghahramani 2002). Though label propagation would appear to be a reasonable approach to our problem, in our case we are using clusters—rather than neighborhoods—over the label space—not the feature space—and all of our items have (noisy) labels. This is partly due to observations about how label sets we have collected appear to cluster in space (Figure 2).

A number of researchers have used **clustering** among related data items to improve the quality of ground-truth labels. For instance, Zhang et al. (2016) show that the latent classes determined by their clustering models are, compared to plurality-based labels, better estimates of the semantics of the data items they study, thus providing support for this approach in the context of supervised learning (which they do not study). McCallum (1999) studies clustering in a semisupervised learning context. In both cases, however, *the clustering is in feature space, not label distribution space*, and is considered part of the learning process. We, on the other hand, use clustering to establish better estimates of ground truth to a secondary, supervised learning step, and we perform cluster in the space of label distributions, not the feature space.

3 Label Distribution Learning on Populations

The *population label distribution learning problem* is to learn to predict the distribution of labels $\mathbf{y}$ among a population of annotators for each test set data item $\mathbf{x}$, given a collection of training data items $(\mathbf{x}_i)_{i\in\{1,\ldots,n\}}$ and a corresponding collection of label distribution *raw estimates* $(\hat{\mathbf{y}}_i)_{i\in\{1,\ldots,n\}}$, based on the normalized *empirical label distributions*, i.e., the distributions of the annotations received for each data item. Note that, here, we assume these distributions are multinomial samples of the underlying population of annotator's *true label distribution* $(\mathbf{y}_i)_{i\in\{1,\ldots,n\}}$, and that the each raw estimate was obtained by randomly choosing an annotator and then asking that annotator to choose a label, then repeating this process m times, where m is a parameter of the sampling process.

One example of a label set that supports this problem definition came from an effort to model Twitter discourse on life trajectories. When inspecting annotators' answers to a question which identifies employment transition events, we observed that when there was disagreement it was often for good reason.

Figure 2 shows the label distributions over the the jobQ3MT+ label set (see more details in Section 4.1). These histograms of labels (one histogram per data item) appear to cluster into approximately eight categories, where the tweets in each seemed to be similarly rated. Group 1 (red) distributions have most of their mass on *Getting hired/job seeking* and *None of the above, but job-related*, with tweets talking about plans to get a job (e.g., *really want a job*, *dont put that on ur resume for a minimum wage job*) or the process of getting a job. Group 2 (cyan) has almost all the mass exclusively on *Getting hired/job seeking* (e.g., *got the job*). Group 3 (brown) clusters around *Complaining about work* and *Going to work*, suggesting a topic about complaining about having to go to work. Group 4 (green) are a set of tweets complaining about work while at work. Groups 5 and 6 (blue and orange) have their peaks on *None of the above, but job-related* and *Not job-related*. Group 6 (where *Not job-related* was more frequent than *None of the above*) were mostly about road work. Group 7 seemed to contain cases where work was mentioned, but not central (e.g., *Today at work I learned about...*) or used "work" or "job" metaphorically, though there exist some clear *None of the above, but job-related* tweets, like *Perks of working overnight: donuts fresh out of the fryer*.

As to why such clustering happens, Zhang et al. (2016), on a different dataset, noticed similar clustering patterns. We note that any k-choice annotation task effectively reduces the full breadth of interpretations encoded in each data item x to one of only k choices; We theorize that the act of annotation reduces not only the interpretive domain of the each data item, but also the social, experiential and cognitive factors, such as disparities in experience and knowledge, that drive annotator disagreement. Thus, the number p of *distinct* ground truth label distributions resulting from any annotation task are also limited, and the set of all annotations for any given data item is (assuming annotators are selected i.i.d. from the population of annotators) a sample from one of the p distinct ground truth distributions. For the sake of brevity we subsequently refer to this tentative explanation as the **clustering theory**.

3.1 Algorithmic Approach

Our approach to label distribution learning on populations consists of two stages. First, we use unsupervised learning to convert the raw label distribution estimates $(\hat{\mathbf{y}}_i)_{i\in\{1,\ldots,n\}}$, into *refined estimates* $(\hat{\mathbf{y}}'_\mathbf{i})_{i\in\{1,\ldots,n\}}$, by aggregating over similarly related data items. Next, we perform supervised learning on the refined label distributions with unstructured text features and conduct comparative experiments. We discuss each stage below.

3.2 Clustering Algorithms for Estimating Ground Truth

The unsupervised learning algorithms we consider here are consistent, to varying degrees, with the clustering theory. The (finite) multinomial mixture model **F** most directly simulates the sampling process according to the cluster theory. It assumes that the empirical label distributions are generated by, (1) drawing a multinomial distribution π according to a Dirichlet prior over p elements (i.e., corresponding to the hypothesized number of true label distributions) $\pi \sim \text{Dir}(p, \gamma = 75)$, where γ is the prior's (symmetric) hyperparameter (and higher numbers tend to produce lower entropy multinomials); (2) drawing multinomial distributions $\phi_1, \ldots \phi_p \sim \text{Dir}(d, \gamma = 0.1)$; (3) for each data item, we (3a) choose $i \sim \pi$ and (3b) m labels according to ϕ_i. Thus, according to the clustering theory the most likely cluster distribution ϕ_j for each data item should be a good estimate of the true label distribution: $\phi_j \approx \mathbf{y}_i$. We use a variational Bayes algorithm[1] to learn the model.

Next come two variants of **F**. The Dirichlet process multinomial mixture model (**DP**) is a non-parametric version of **F**. Instead of choosing p multinomial models from a Dirichlet prior before generating the data, it starts with two multinomial models $\phi_1, \phi_2 \sim \text{Dir}(d, 0.1)$. Then, for each new data item it draws from the current set of multinomial models in approximate proportion to the number of times each has been previously drawn OR draws a new multinomial model (with weight proportional to $\gamma = 50$). We use a variational Bayes algorithm to learn this model. The main purpose for including it here is to test whether in this setting nonparametric methods outperform parametric ones using standard model-selection criteria.

M is a multinomial mixture model without Dirichlet priors. This rather simple model can be learned using EM, however it lacks the regularization and adaptability that the Dirichlet priors provide. We expect this model to underperform the others.

In contrast to the previous models, we chose the Gaussian mixture model **G** as a weak alternate hypothesis of sorts. Rather than simulate the sampling process, as the multinomial distributions do, these distributions capture the variance in a population of samples. Additionally, it captures covariance between the labels; these should be close to zero in single label settings (or settings where the vast majority of annotators provide only one label per item). We use EM[2] to learn this model.

Finally, **L** is latent Dirichlet allocation[3] (Blei, Ng, and Jordan 2003). Though LDA is not a proper clustering model, we can obtain cluster-like latent classes from it. In terms of **F**, rather than choosing a single class selection distribution π for all data items, it chooses a new one π_i for each item i and for each label chooses a new distribution in $\{\phi_1, \ldots, \phi_p\}$ according to π_i. Thus, each instance of the labels for each item i from LDA represent a true mixture of all of the generating distributions, and is therefore not a proper clustering model (in contrast to the other models, where each instance of labels comes from one generating distribution only, although different instances may use different generators). Nonetheless, we can "assign" to i the most likely ϕ_j according to π_i.

[1] Adapted from https://github.com/bnpy/bnpy.

[2] http://scikit-learn.org/stable/modules/mixture.html

[3] Based on https://radimrehurek.com/gensim/

3.3 Supervised Learning for Predicting Label Distributions

We train supervised-learning-based classifiers using refined label distributions obtained from the various unsupervised learning algorithms described above. We retain the most common 20,000 words in the test set and pad the sentence with up to 1,000 tokens in the text pre-processing step, then embed each word into a 100-dimension vector using the GloVe 2B-tweet corpus (Pennington, Socher, and Manning 2014).

We consider two neural network models. One ("**CNN**") is based on a 1D convolutional neural network, designed for sentence or tweet classification (Kim 2014), with three max pool/convolution layers, followed by a dropout and a softmax layer. The other ("**LSTM**") is an encoder-decoder sequence-to-sequence model using recurrent neural networks . The encoder outputs a fixed-length encoding of the input text, and the decoder predicts the output sequence.

For both types of models, we use *softmax*: $\frac{exp(z_i)}{\Sigma_t exp(z_t)}$, to transform the output of the penultimate layer $\mathbf{z}$ into a probability distribution. We use *Kullback-Leibler (KL) divergence*, a standard measure of the difference between the "true" (in our case the refined estimate) probability distribution $\hat{\mathbf{y}}'$ and a predicted estimator $\tilde{y}$: $\text{D}_{KL}(\hat{y}', \tilde{y}) = \sum_i P(\hat{y}' = i) \frac{\log P(\hat{y}'=i)}{\log P(\tilde{y}=i)}$, as the loss function for backpropagation for this is a principled choice which would approximate the full probability distributions (Vieira 2014), with the *Adam* optimizer (Kingma and Ba 2014). We train with a batch size of 32 and 25 epochs.

4 Experiments

4.1 Data and Labels

We consider two corpora, each consisting of 2,000 tweets, one related to work (mentioned in Section 3), the other to suicide. Our institutional review board determined that this research fell into the exempt class. To privatize the data we replaced all mentions of usernames with "@SOMEONE" and URLs with "http://URL," and adhered to Twitter's developer policy (Twitter 2018). Table 1 gives basic properties of the labels we collected for these two corpora.

Job-related We introduced the job dataset in Section 3. It contains 2,000 tweets about work that were extracted by a publicly available library (Liu et al. 2016). We asked five crowdworkers each from Figure Eight (**FE**, 2019) and Amazon Mechanical Turk (**MT**, 2019) to answer three questions about each tweet. Figure 3 shows the three questions we asked and their corresponding selections of labels. We denote these label sets *jobQ1/2/3*. To provide some insight into how performance might change with more labels from a more diverse population of labelers and labeling platforms,

we first consider FE and MT as two separate label sets, then combine them into a single label set (denoted BOTH).

For each question, we then run experiments on two different train/dev/test splits. We first consider a 1000/500/500 split on each of the label sets: Q1, Q2, and Q3 (which we call the **Broad** split). Next, to get a more accurate ground-truth estimate for testing we randomly selected 50 tweets from our dataset and asked 50 additional MT crowdworkers to label them. We denote these label sets *jobQ1/2/3MT+* and create 1500/450/50 splits (called the **Deep** splits), where the training and development sets are from the BOTH label sets (minus the *jobQ1/Q2/Q3MT+* label set items) and the test sets are from *jobQ1/Q2/Q3MT+*, respectively.

Q1. Which of the following items could best describe the point of view of job /employment-related information in the target tweet?

- ○ 1st person
- ○ 2nd person
- ○ 3rd person
- ○ Unclear
- ○ Not job-related

Q2. Which of the following items could best describe the employment status of the subject in the tweet?

- ○ Employed
- ○ Not Employed
- ○ Not in Labor Force
- ○ Unclear
- ○ Not job-related

Q3. Does the subject specifically mention any job/employment transition event in the tweet? (Choose all that apply)

- ☐ 1 Getting hired/job seeking
- ☐ 2 Getting Fired
- ☐ 3 Quitting a job
- ☐ 4 Losing job some other way
- ☐ 5 Getting promoted/raised
- ☐ 6 Getting cut in hours
- ☐ 7 Complaining about work
- ☐ 8 Offering support
- ☐ 9 Going to work
- ☐ 10 Coming home from work
- ☐ 11 None of the above, but job-related
- ☐ 12 Not job-related

Figure 3: The job-related annotation tasks contain these three questions and corresponding choices. The answers for Q3 are the columns in each of the histograms in Figure 2.

Suicide-related The *Suicide* tweet label set was obtained directly from (Liu et al. 2017). It contains for each data item labels from five Figure Eight crowdworkers and up to two experts in suicide prevention. Each tweet was labeled as one of the following: Ⓐ *Suicidal thoughts*, Ⓑ *Supportive messages or helpful information*, Ⓒ *Reaction to suicide news/movie/music* and Ⓓ *Others*. We use a 1000/500/500 train/dev/test split.

4.2 Clustering Experiments for Ground Truth Estimation

Model Selection For those clustering models requiring p as a hyperparameter, we test values for $p \in [d/2, 2d]$, where d is the number of label choices. As the estimators for these models are stochastic and/or sensitive to initial conditions, for every model and every set of hyperparameters we ran 100 trials on the training/dev set and picked the model with the highest estimated likelihood. Table 2 shows the number of clusters selected on each of the two training splits on each label set and for **DP** the number of clusters the algorithm generated.

Label Set	#Items	#Choices /item	#Workers	#Labels	Density	MVTD	RMSD
jobQ1FE	2,000	5	171	10,000	5.00	0.37	0.21
jobQ1MT	2,000	5	1,014	12,202	6.10	0.17	0.10
jobQ1BOTH	2,000	5	1,185	22,202	11.10	0.29	0.16
jobQ1MT+	50	5	249	2,969	59.38	0.43	0.22
jobQ2FE	2,000	5	171	10,000	5.00	0.28	0.16
jobQ2MT	2,000	5	1,014	12,202	6.10	0.15	0.09
jobQ2BOTH	2,000	5	1,185	22,202	11.10	0.23	0.13
jobQ2MT+	50	5	249	2,969	59.38	0.34	0.19
jobQ3FE	2,000	12	171	10,967	5.48	0.45	0.16
jobQ3MT	2,000	12	1,014	12,900	6.45	0.28	0.10
jobQ3BOTH	2,000	12	1,185	23,867	11.93	0.40	0.14
jobQ3MT+	50	12	249	3,196	63.92	0.41	0.14
Suicide	2,000	4	124	13,175	6.59	0.27	0.17

Table 1: Basic properties of our label sets. For the job-related data set with three questions *jobQ1/2/3*, *FE* and *MT* represent the labels from the platforms Figure Eight and Amazon Mechanical Turk respectively. *BOTH* combines both FE and MT labels. **Density** is the average number of labels per data item. **MVTD** (majority-voted-true-class deviation) and **RMSD** (root-mean-square deviation) describe inter-rater reliability across all the tasks and estimate the variety and divergence of different label sets, motivated by the literature on scale and outlier description (Hyndman and Koehler 2006; Pontius, Thontteh, and Chen 2008; Willmott and Matsuura 2006). MVTD is the average deviation of the majority-voted label over all data items: MVTD = $1 - \sum_{i=1}^{n} \max_j\{\hat{y}_{ij}\}/n$. RMSD is the L2 deviation from the average label distribution: RMSD = $\sum_{i=1}^{n} \sqrt{(\hat{\mathbf{y}}_i - \overline{y})^T(\hat{\mathbf{y}}_i - \overline{y})}/n$, where $\overline{y}$ is the average label distribution over all data.

	Broad split					Deep split				
Dataset	**M**	**G**	**L**	**F**	**DP**	**M**	**G**	**L**	**F**	**DP**
jobQ1FE	10	4	9	3	4	11	11	9	3	4
jobQ1MT	11	4	11	8	10	2	2	11	9	11
jobQ1BOTH	11	2	2	6	8	2	2	11	7	8
jobQ2FE	11	3	10	3	4	11	2	10	3	4
jobQ2MT	2	4	11	7	9	2	2	11	7	10
jobQ2BOTH	2	2	11	5	7	2	2	8	5	7
jobQ3FE	19	5	18	6	7	19	10	19	7	7
jobQ3MT	5	5	14	17	20	5	19	15	17	26
jobQ3BOTH	5	15	18	13	16	5	17	11	17	17
Suicide	8	2	7	4	5	-	-	-	-	-

Table 2: The optimal label aggregation models on each label set using two splits (*Broad* and *Deep*) are achieved with the presented number of clusters (p).

Evaluation For the model M produced by each unsupervised learning algorithm and each data item i in the test set, we determine the most likely cluster j for i's empirical label distribution ϕ_j: $\arg\max_j P(\hat{y}_i \sim \phi_j \mid M)$. We then compute the KL divergence between the empirical label distribution $\hat{y}_i$ and the cluster distribution ϕ_j.

Table 3 shows that the multinomial mixture models (**M/F/DP**) generally outperformed **G**, as we expected. The crowdsourced sample sizes of 5–10 labels we used for each training item are typical of crowdsourced supervised learning label sets, and the differences between **G** and the other cluster models appear to be substantial at this scale. The success of **L** on a number of label sets surprised us, considering

that we only use the mostly likely cluster for each data item which was trained on a mixture of clusters. Finally, **F** outperforms the other models on all of the sets having at least ten annotations per item, and shows the most improvement from the FE/MT (which had five annotations per item) to the BOTH (with ten annotations per item) label sets.

	Broad split					Deep split				
KL	**M**	**G**	**L**	**F**	**DP**	**M**	**G**	**L**	**F**	**DP**
jobQ1FE	0.35	0.53	0.23	0.39	0.39	0.30	0.57	0.24	0.37	0.39
jobQ1MT	0.19	0.68	0.18	0.13	0.15	0.20	0.39	0.07	0.09	0.10
jobQ1BOTH	0.20	0.46	0.40	0.19	0.19	0.21	0.38	0.06	0.06	0.07
jobQ2FE	0.26	0.54	0.19	0.32	0.32	0.24	0.65	0.20	0.28	0.28
jobQ2MT	0.36	0.74	0.15	0.10	0.10	0.26	0.50	0.09	0.11	0.13
jobQ2BOTH	0.28	0.51	0.17	0.16	0.16	0.25	0.48	0.09	0.08	0.08
jobQ3FE	0.51	1.00	0.52	0.59	0.64	0.29	0.97	0.27	0.41	0.41
jobQ3MT	0.50	1.15	0.33	0.26	0.29	0.20	0.51	0.17	0.28	0.21
jobQ3BOTH	0.45	0.82	0.35	0.32	0.33	0.18	0.64	0.18	0.12	0.13
Suicide	0.22	0.57	0.20	0.22	0.22	-	-	-	-	-
Average	0.29	0.59	0.28	0.22	0.23	0.21	0.50	0.11	0.09	0.09
Std dev	0.10	0.14	0.10	0.06	0.06	0.03	0.11	0.05	0.02	0.03

Table 3: KL divergence based on the chosen label clustering models in Table 2. Average and standard deviation are based on the KL divergence scores of the dark gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide). The *lowest* KL is highlighted in light gray for each split.

Table 3 also shows the average and standard deviation of the KL divergence scores on the four independent label sets (i.e., BOTH comprises FE and MT) jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide (highlighted in gray). These statistics indicate that **F** outperforms the other models across different thematic label sets in its capability and stability, **DP** is second, and, as we expected, **G** comes last.

Q3 differs from Q1 and Q2 in allowing annotators to choose multiple labels. Ideally, then the ideal representation for the annotations (where each *annotation* is the set of labels provided by one annotator for one data item) of Q3 would be over the *power set* of possible labels. However, Table 4 shows that fewer than 10% of the annotations we received had selected more than one label. To simplify our analysis, we thus treat multiple labels from the same annotator as if each came from its own, independent annotator (for example, an annotation with three labels provided is treated as three separate annotations.).

	#labels/worker/item				
Label Set	**1**	**2**	**3**	**4**	**5+**
jobQ3FE	10,000	722	176	53	16
jobQ3MT	12,202	628	58	11	1
jobQ3MT+	2,969	193	32	2	0

Table 4: Counts of worker-item pairs, grouped by #labels per worker per item.

4.3 Supervised Learning Experiments

We then trained the two supervised learning algorithms described in Section 3.3 on our training datasets' texts, using in turn each of the unsupervised learning methods described previously to provide *refined label distribution estimates* ($\hat{\mathbf{y}}'_i$) as the learning goal. We compared their performances to those of three common baseline strategies for resolving (or not) label disagreement.

- Majority (**Maj**) takes the final label to be $\hat{y}'_i = \arg\max_{j\in\{1,\ldots,d\}} \{\hat{y}_{ij}\}$.
- Repeated (**Rept**) duplicates each data instance once for every annotation it receives and pairs the replicated instance with that label.
- Probability (**Prob**) is the raw label distribution estimates $(\hat{\mathbf{y}}'_i) = (\hat{\mathbf{y}}_i)$. (This is the baseline LDL approach.)

Evaluation We measure the **KL divergence** between the classifier ($\tilde{\mathbf{y}}_i$) and cluster-or-baseline-method ($\hat{\mathbf{y}}'_i$) -based label distributions. (Note that Maj and Rept both associate each data item, by eliminating labels or creating copies of the data items, exactly one label. For the purpose of computing KL divergence we regard this as a distribution where the entire probability mass is on one label.) We also measure **Accuracy**, i.e., the percentage of times $\arg\max_j \tilde{y}_{ij}$ matches $\arg\max_j \hat{y}'_{ij}$ in the test set. Accuracy is often used in nondistributional classification problems. We use it here to shed further light into the differences between distributional and nondistributional problems. In particular, we might expect that nondistributional models might outperform label distribution models with respect to accuracy, even as they underperform with respect to KL divergence.

Results Tables 5 and 6 show the KL divergence and accuracy metrics for CNN/LSTM text classifiers built with different label aggregation strategies in two split modes.

Starting with the KL divergence results, on the Broad split tests, CNNs trained and tested on **L** outperform other clustering and non-clustering approaches most of the time for both job and suicide discourse themes. For LSTMs, we can also observe that clustering approaches achieved better results more often on different label sets than non-clustering methods. Almost none of CNNs or LSTMs trained on any baseline label reduction strategy can compete.

By contrast, the results of the Deep split KL divergence tests (Table 6) are not as conclusive, and this could be due to there being fewer data items in the Deep split test set. But even so, clustering strategies again perform better in more cases than the baselines.

Tables 5 and 6 show that, for both the CNN and LSTM classifiers and both split modes, the highest accuracies often come from the clustering methods. They outperform non-clustering methods by more than 10% on average, which appears substantial. For those label sets whose accuracy based on clustering strategies do not rank 1st, non-clustering methods win only by a slim or zero margins.

Together, the results for different label sets and split modes reveal several interesting patterns. First, the cluster-based models tend to outperform the baseline methods in terms of either KL divergence or accuracy as reported. This supports the feasibility of our clustering strategy for label distribution learning on subjective problems with annotator

CNN	KL divergence								Accuracy							
	Maj	Rept	Prob	M	G	L	F	DP	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	2.98	0.79	0.91	0.12	0.74	0.47	0.18	0.19	0.73	0.53	0.72	0.78	0.95	0.58	0.64	0.58
jobQ1MT	2.03	0.80	0.72	0.65	1.05	0.52	1.02	1.00	0.80	0.72	0.79	0.56	0.67	0.76	0.54	0.56
jobQ1BOTH	2.38	0.45	0.48	0.36	0.38	0.27	0.40	0.38	0.82	0.64	0.81	0.57	0.76	0.76	0.65	0.64
jobQ2FE	2.29	0.91	0.79	0.21	0.78	0.13	0.31	0.28	0.73	0.63	0.79	0.71	0.62	0.94	0.59	0.64
jobQ2MT	2.10	0.80	0.78	0.81	0.98	0.67	1.04	0.96	0.73	0.68	0.73	0.48	0.55	0.71	0.53	0.52
jobQ2BOTH	2.12	0.49	0.47	0.48	0.48	0.37	0.51	0.52	0.76	0.65	0.76	0.63	0.58	0.71	0.54	0.56
jobQ3FE	4.20	1.66	1.14	0.31	0.68	0.66	0.42	0.36	0.36	0.31	0.41	0.47	0.32	0.45	0.42	0.46
jobQ3MT	3.18	2.24	1.05	1.04	1.32	0.54	1.12	1.12	0.53	0.45	0.51	0.26	0.28	0.49	0.28	0.28
jobQ3BOTH	3.38	1.40	0.77	0.62	0.49	0.62	0.71	0.70	0.48	0.42	0.53	0.31	0.62	0.46	0.25	0.21
Suicide	2.16	1.40	0.45	0.69	13.62	0.33	0.53	0.49	0.81	0.65	0.78	0.18	1.00	0.76	0.37	0.39
Average	2.51	0.94	0.54	0.54	3.74	0.40	0.54	0.52	0.72	0.59	0.72	0.42	0.74	0.67	0.45	0.45
Std dev	0.51	0.47	0.13	0.13	5.70	0.13	0.11	0.11	0.14	0.10	0.11	0.18	0.16	0.12	0.15	0.17
LSTM	Maj	Rept	Prob	M	G	L	F	DP	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.80	0.91	1.12	0.49	0.66	0.63	0.33	0.53	0.84	0.75	0.87	0.89	0.99	0.76	0.83	0.84
jobQ1MT	1.09	1.16	1.15	1.37	1.49	1.40	1.07	0.62	0.86	0.82	0.85	0.81	0.87	0.83	0.80	0.81
jobQ1BOTH	0.75	0.80	0.54	1.12	0.45	0.83	0.52	1.09	0.88	0.79	0.86	0.81	0.89	0.82	0.82	0.82
jobQ2FE	1.25	1.12	1.20	0.79	0.94	1.14	1.14	0.54	0.85	0.78	0.87	0.85	0.82	0.97	0.84	0.82
jobQ2MT	1.88	1.07	1.48	0.92	1.52	1.24	1.71	1.25	0.84	0.81	0.82	0.78	0.83	0.81	0.79	0.80
jobQ2BOTH	0.86	0.78	1.68	1.50	0.67	0.93	0.89	0.73	0.86	0.80	0.84	0.86	0.81	0.81	0.80	0.83
jobQ3FE	1.79	1.68	1.54	0.93	1.13	1.14	1.01	0.92	0.64	0.64	0.62	0.65	0.72	0.71	0.65	0.62
jobQ3MT	2.08	1.65	1.86	1.81	1.73	1.18	1.88	1.58	0.69	0.70	0.63	0.63	0.59	0.50	0.66	0.67
jobQ3BOTH	1.26	1.46	1.99	1.46	1.10	1.38	1.42	1.40	0.70	0.68	0.63	0.63	0.86	0.61	0.61	0.66
Suicide	1.14	0.74	0.85	0.67	13.97	0.91	0.68	0.85	0.74	0.72	0.74	0.73	0.50	0.72	0.71	0.71
Average	1.00	0.95	1.27	1.19	4.05	1.01	0.88	1.02	0.80	0.75	0.77	0.76	0.77	0.74	0.74	0.76
Std dev	0.21	0.30	0.59	0.33	5.73	0.22	0.34	0.26	0.08	0.05	0.09	0.09	0.16	0.08	0.08	0.07

Table 5: KL divergence and accuracy of the **Broad** split. Average and standard deviation are based on the dark gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide). The *lowest* KL and *highest* accuracy are highlighted in light gray.

CNN	KL divergence								Accuracy							
	Maj	Rept	Prob	M	G	L	F	DP	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	3.09	0.77	0.90	0.13	0.69	0.39	0.09	0.16	0.62	0.47	0.58	0.78	0.80	0.54	0.82	0.72
jobQ1MT	2.94	0.47	0.54	0.64	1.08	0.47	1.22	1.05	0.72	0.53	0.70	0.58	0.66	0.72	0.56	0.58
jobQ1BOTH	2.90	0.34	0.24	0.39	0.43	0.38	0.33	0.35	0.72	0.51	0.70	0.62	0.90	0.60	0.62	0.60
jobQ2FE	3.07	0.57	0.65	0.18	0.56	0.49	0.21	0.31	0.60	0.53	0.52	0.76	0.82	0.48	0.50	0.60
jobQ2MT	1.90	0.50	0.58	0.77	0.68	0.76	0.74	1.07	0.72	0.57	0.70	0.58	0.64	0.66	0.64	0.44
jobQ2BOTH	2.90	0.27	0.28	0.52	0.37	0.35	0.50	0.58	0.72	0.54	0.76	0.54	0.56	0.68	0.64	0.54
jobQ3FE	3.71	1.45	1.00	0.34	0.63	0.65	0.53	0.43	0.46	0.40	0.48	0.16	0.30	0.50	0.14	0.40
jobQ3MT	3.95	1.98	0.77	1.13	1.21	1.20	1.26	1.24	0.54	0.43	0.54	0.14	0.24	0.48	0.30	0.18
jobQ3BOTH	3.33	1.13	0.63	0.76	0.67	0.49	0.71	0.73	0.62	0.46	0.48	0.20	0.40	0.56	0.16	0.24
Average	3.04	0.58	0.38	0.56	0.49	0.41	0.51	0.55	0.69	0.50	0.65	0.45	0.62	0.61	0.47	0.46
Std dev	0.20	0.39	0.18	0.15	0.13	0.06	0.16	0.16	0.05	0.03	0.12	0.18	0.21	0.05	0.22	0.16
LSTM	Maj	Rept	Prob	M	G	L	F	DP	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.94	0.85	0.76	0.52	0.89	0.61	0.79	0.65	0.74	0.70	0.74	0.92	0.93	0.67	0.92	0.91
jobQ1MT	0.78	0.53	0.52	0.88	1.08	1.10	1.80	1.38	0.71	0.71	0.71	0.79	0.87	0.70	0.82	0.82
jobQ1BOTH	0.39	0.65	0.70	0.63	0.72	0.54	0.86	0.64	0.70	0.71	0.70	0.81	0.98	0.69	0.84	0.85
jobQ2FE	0.99	1.16	1.04	1.37	0.57	0.90	0.98	0.88	0.77	0.72	0.77	0.87	0.91	0.75	0.85	0.86
jobQ2MT	0.83	0.94	0.77	0.96	1.44	1.32	1.15	1.13	0.69	0.69	0.69	0.76	0.83	0.66	0.81	0.81
jobQ2BOTH	0.88	0.63	0.73	0.67	1.30	0.94	1.31	1.15	0.69	0.69	0.69	0.86	0.85	0.64	0.80	0.83
jobQ3FE	1.97	1.55	1.40	0.84	0.90	1.71	0.92	1.04	0.62	0.65	0.67	0.66	0.83	0.70	0.64	0.59
jobQ3MT	1.51	1.38	1.44	1.65	2.01	1.99	1.63	1.67	0.68	0.61	0.62	0.64	0.59	0.62	0.67	0.60
jobQ3BOTH	1.41	1.28	1.26	1.43	1.06	1.54	1.29	1.46	0.62	0.61	0.68	0.64	0.71	0.67	0.68	0.58
Average	0.89	0.85	0.90	0.91	1.03	1.01	1.15	1.08	0.67	0.67	0.69	0.77	0.85	0.67	0.77	0.75
Std dev	0.42	0.30	0.26	0.37	0.24	0.41	0.21	0.34	0.04	0.04	0.01	0.09	0.11	0.02	0.07	0.12

Table 6: KL divergence and accuracy of the **Deep** split. Average and standard deviation are based on the dark gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, and jobQ3BOTH). The *lowest* KL and *highest* accuracy are highlighted in light gray.

disagreement. On the other hand, for conventional (i.e., non-distributional) classification problems, baseline methods can be sufficient, as shown in our experiment results. The advantages of clustering, in terms of KL divergence, is less stark in the Deep compared to the Broad splits, but clustering still seems to outperform baselines on the jobQ3 label set, which has the largest label space and is where pooling and other label conservation methods are most needed.

5 Discussion

Our results provides evidence—both for and against—that clustering is a feasible strategy to improve performance of label distribution learning in certain settings, such as when each label distribution represents a population estimate based on a (micro) sample, and the data falls into a small number of semantic equivalence classes (relative to the learning task). Yet, why this is so is still not clear; our results shed little light on the validity of the clustering theory.

They also raise methodological issues. We expect that the methods introduced here for testing performance will provide helpful baselines for the development of newer methods tailored specifically toward settings where ground truth depends on a small number of samples per data item.

One methodological issue we grappled with was whether to measure the performances of the supervised models against the empirical ($\hat{y}$) or refined ($\hat{y}'$) label distributions. Standard practice is to test supervised learning on the patterns they are fed (i.e., the refined labels in our case). But in our case the conventional machine learning algorithms are only the last half of a larger pipeline that has essentially an unsupervised front end, and which takes the empirical labels as input. We tried both approaches, but here, for space purposes and because we found our results more interesting in this direction, we report on only the predictions against $\hat{y}'$. The biggest worry in doing so is that, because pooling labels via a small number of clusters greatly reduces diversity in the label distributions, there is less likelihood of error, which would seem to make predictions artificially easier against $\hat{y}'$ than the empirical distributions $\hat{y}$. On the other hand, since these clusters are based on labels, the larger the clusters the greater the likelihood that items with inconsistent features are assigned to the same cluster, and this would lead to less accurate predictions from the supervised models.

We have been deliberately vague about what "population of labelers" means. This study was motivated by our work with microtask crowdsourcing sites like Amazon Mechanical Turk and Figure Eight, in which case our labels can be taken as collection of (micro) samples of the population of workers on whichever sites are used for whatever interval of time the requested labeling task is posted. Studies exist on the demographics of these sites. Some sites (like Figure Eight in our study) provide some demographic information on the responders to each microtask request.

We have not yet modeled user behavior, though this is a well-established approach for aggregating labels from multiple annotators. We did, in fact, run experiments using Dawid and Skene's class annotator-based model (Dawid and Skene 1979), which is largely based on using behavior. However, as it is designed for conventional, non-distributional supervised learning and did not perform well, we did not report those results here. Another complication is that most of our annotators labeled only ten data items each, so we would be tempted to used clustering to group users in much the same way we used it here to group data items.

Another limitation was that we did not investigate in-depth the causes of inter-annotator disagreement, such as data encoding errors and communication ambiguities (Zhu and Wu 2004; Angluin and Laird 1988; Brodley and Friedl 1999), lack of sufficient information (Hickey 1996; Brodley and Friedl 1999; Brazdil and Clark 1990), and unreliable annotators and their bias (Hickey 1996), nor did we attempt to resolve disagreement through follow-up discussions with the annotators, as is common in many grounded theory studies.

6 Conclusion

We study the important problem of predicting the distributions of population beliefs using both unsupervised and supervised learning methods. We test different strategies for clustering data items to obtain aggregated label distributions. We then build supervised CNN/LSTM classifiers using the predicted distributions and compared the performance with common baseline label reduction strategies. Our results from both unsupervised and supervised experiments show that it is feasible to predict probability distributions over labels at the population level. Clustering labels, in general, boosts the label distribution learning by aggregating data items with similar semantics and population beliefs. We believe our study is an pioneering exploration of disagreement on linguistic data from social media and further helps future intelligent agents understand the diversity of beliefs in society.

7 Acknowledgments

We thank the anonymous reviewers for their helpful feedback and suggestions. Many thanks to Tharindu Cyril Weerasooriya, Lingjia Deng, and Chen Chen for their contributions and conversations.

References

Angluin, D., and Laird, P. 1988. Learning from noisy examples. *Machine Learning* 2(4):343–370.

Aroyo, L., and Welty, C. 2014. The three sides of crowdtruth. *Journal of Human Computation* 1:31–34.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Brazdil, P., and Clark, P. 1990. Learning from imperfect data. In *Machine Learning, Meta-Reasoning and Logics*. Springer. 207–232.

Brodley, C. E., and Friedl, M. A. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research* 11:131–167.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk.

In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, 286–295. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, N.-C.; Drouhard, M.; Kocielnik, R.; Suh, J.; and Aragon, C. R. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems* 8(2).

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

FE. 2019. Figure Eight. https://www.figure-eight.com/. (Accessed June 5, 2019).

Gao, B.-B.; Xing, C.; Xie, C.-W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26(6):2825–2838.

Geng, X., and Hou, P. 2015. Pre-release prediction of crowd opinion on movies by label distribution learning. In *IJCAI*, 3511–3517.

Geng, X.; Wang, Q.; and Xia, Y. 2014. Facial age estimation by adaptive label distribution learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 4465–4470. IEEE.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748.

Hickey, R. J. 1996. Noise modelling and evaluating learning from examples. *Artificial Intelligence* 82(1):157–179.

Hyndman, R. J., and Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22(4):679–688.

Jia, X.; Li, W.; Liu, J.; and Zhang, Y. 2018. Label distribution learning by exploiting label correlations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, T.; Homan, C. M.; Alm, C. O.; White, A. M.; Lytle, M. C.; and Kautz, H. A. 2016. Understanding discourse on work and job-related well-being in public social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1044–1053. Berlin, Germany: Association for Computational Linguistics.

Liu, T.; Cheng, Q.; Homan, C.; and Silenzio, V. 2017. Learning from various labeling strategies for suicide-related messages on social media: An experimental study. In *The workshop on Mining Online Health Reports of the 10th ACM Conference on Web Search and Data Mining*.

McCallum, A. 1999. Multi-label text classification with a mixture model trained by em. In *AAAI workshop on Text Learning*, 1–7.

MT. 2019. Amazon Mechanical Turk. https://www.mturk.com/. (Accessed June 5, 2019).

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pontius, R. G.; Thontteh, O.; and Chen, H. 2008. Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* 15(2):111–142.

Ren, Y., and Geng, X. 2017. Sense beauty by label distribution learning. In *Proc, IJCAI*, 2648–2654.

Schaekermann, M.; Law, E.; Williams, A. C.; and Callaghan, W. 2016. Resolvable vs. irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth. In *Proceedings of the 1st Workshop on Human-Centered Machine Learning at SIGCHI*.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622. ACM.

Shirani, A.; Dernoncourt, F.; Asente, P.; Lipka, N.; Kim, S.; Echevarria, J.; and Solorio, T. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 1167–1172.

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 conference on empirical methods in natural language processing*, 254–263. Association for Computational Linguistics.

Twitter. 2018. Twitter Developer Terms. https://developer.twitter.com/en/developer-terms/agreement-and-policy. Online, Accessed June 5, 2019.

Vieira, T. 2014. Kl-divergence as an objective function. https://timvieira.github.io/blog/post/2014/10/06/kl-divergence-as-an-objective-function/. Online, Accessed August 14, 2019.

Willmott, C. J., and Matsuura, K. 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science* 20(1):89–102.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26(8):1819–1837.

Zhang, J.; Sheng, V. S.; Wu, J.; and Wu, X. 2016. Multi-class ground truth inference in crowdsourcing with clustering. *IEEE Trans. Knowl. Data Eng.* 28(4):1080–1085.

Zhu, X., and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation.

Zhu, X., and Wu, X. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review* 22(3):177–210.