Rationale-aware Label Aggregation in Crowdsourcing

TOMOYA NISHIO, University of Tsukuba, Japan
HIROYOSHI ITO, University of Tsukuba, Japan
TAKUMI TAMURA, University of Tsukuba, Japan
ATSUYUKI MORISHIMA, University of Tsukuba, Japan

In recent years, crowdsourcing, which seeks to solve problems by outsourcing tasks to a large, unspecified group of people, has garnered significant attention. Due to the nature of delegating tasks to a diverse and unknown workforce, quality control remains a critical challenge in crowdsourcing. A common approach to quality control is to assign multiple workers to the same task and aggregate their responses. Additionally, quality control methods that utilize the rationale behind workers' answers have been shown to be effective. This paper proposes RAAG (Rationale-aware AGgregation), an aggregation method for quality control in crowdsourcing that incorporates workers' response rationales. RAAG models the process by which workers provide both labels and rationales as a probabilistic generative model. Using this model, we estimate the expected true label of a task as well as latent variables. Specifically, we employ the Expectation-Maximization (EM) algorithm, where the E-step estimates the expected true label, and the M-step updates the latent variables. Experimental evaluations using both synthetic and real-world datasets demonstrate that RAAG achieves higher aggregation quality compared to the DS method. Notably, RAAG exhibits significant improvements in accuracy under challenging conditions where existing methods struggle, such as scenarios with low redundancy, a limited number of responses per worker, and the presence of spam workers.

Additional Key Words and Phrases: Human Computation, Crowdsourcing, algorithm

1 Introduction

In recent years, there has been growing interest in crowdsourcing, which aims to solve problems by delegating tasks to an unspecified number of workers. Specifically, it is used for tasks such as data annotation in information retrieval [1, 2, 11], natural language processing [6, 17, 20, 21], and computer vision [16, 18, 30], as well as in services like ride-hailing [27] and delivery [26]. Since crowdsourcing relies on delegating tasks to an unspecified number of workers, the abilities and motivations of the workers can be unstable, making quality control crucial [15].

In particular, for multi-class classification problems in annotation tasks, it is common to assign multiple workers to the same data and aggregate their results as part of quality control. The simplest aggregation method is majority voting, and a standard method for multi-class classification problems is the Dawid-Skene (DS) method [4].

Aggregation methods such as DS assume that the data obtained through crowdsourcing platforms only includes the final answers. Therefore, the process behind a worker's response is unclear, and the explainability of the responses is low. However, studies have reported that prompting workers to provide rationales behind their answers improves both the explainability and reliability of the responses [19], indicating that considering the rationales behind answers is effective for quality control.

Authors' Contact Information: Tomoya Nishio, University of Tsukuba, Tsukuba, Japan, s2521713@u.tsukuba.ac.jp; Hiroyoshi Ito, University of Tsukuba, Tsukuba, Japan, ito@slis.tsukuba.ac.jp; Takumi Tamura, University of Tsukuba, Tsukuba, Japan, tamura.takumi.ap@alumni.tsukuba.ac.jp; Atsuyuki Morishima, University of Tsukuba, Tsukuba, Japan, mori@slis.tsukuba.ac.jp.

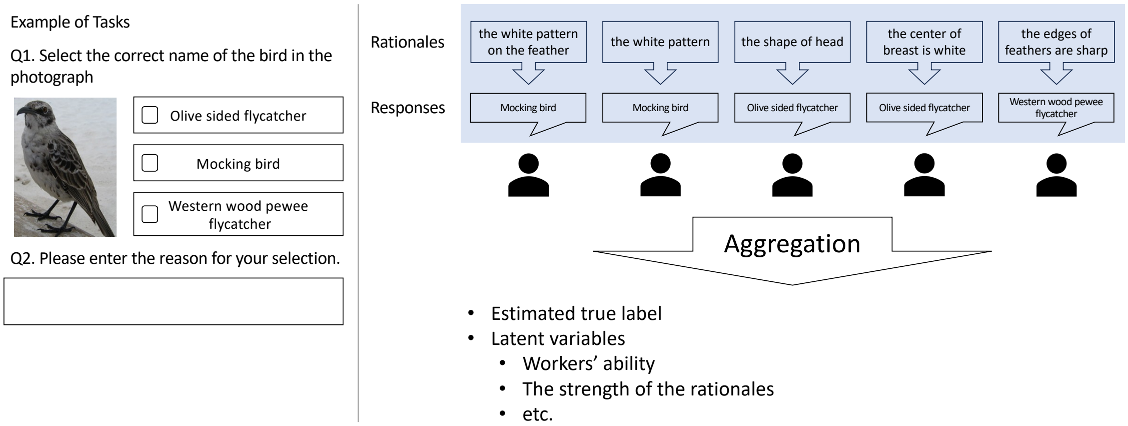

Example of Tasks
Q1. Select the correct name of the bird in the photograph
Olive sided flycatcher
Mocking bird
Western wood pewee flycatcher
Q2. Please enter the reason for your selection.
Rationales
Responses
the white pattern on the feather
the white pattern
the shape of head
the center of breast is white
the edges of feathers are sharp
Mocking bird
Mocking bird
Olive sided flycatcher
Olive sided flycatcher
Western wood pewee flycatcher
Aggregation
Estimated true label
Latent variables
Workers' ability
The strength of the rationales
etc.

Fig. 1. Overview

While there are studies that use rationales behind answers during the aggregation process or in task design [10, 13, 19], no research has directly applied the rationales behind answers to the aggregation process. Therefore, this paper investigates the direct use of rationales behind answers in the DS method and sets the following research questions, which will be clarified through experimentation.

RQ1: *Does an aggregation method using workers' rationales improve the quality of the aggregation results?*

RQ2: *Under what conditions does the method improve the quality of the aggregation results?*

Figure 1 illustrates an example in our problem setting. We assume that we have a set of multiple classification tasks. In each task, the worker gives not only a class, but also his or her rationale behind the decision. The rationale can be given in a paragraph (such as "the white pattern on the feather") or in a URL that contains evidence to support the decision. We want to aggregate the results from more than one worker to obtain the aggregated result, considering the rationales for the classification results.

To answer the research questions, we devised an aggregation method named RAAG (Rationale-Aware AGregation), which models the process by which workers provide both labels and rationales as a probabilistic generative model in crowdsourcing. By using this model[1], RAAG estimates the expected true label of a task, as well as latent variables such as the workers' abilities and the strength of their rationales. In the actual estimation, we use the EM algorithm to estimate the expected true label in the E-step and the latent variables in the M-step.

Then, we conducted simulation experiments with synthetic data and experiments with real-world data, followed by an analysis of the aggregation results.

The results showed that RAAG aggregates labels with higher accuracy compared to existing methods. In particular, it remarkably improved accuracy over existing methods in situations with low redundancy, few responses per worker, and the presence of spam workers. Furthermore, experiments with real-world data also showed that RAAG achieved superior aggregation quality compared to existing methods.

The contributions of this paper are as follows:

- Showed that using rationales behind answers in label aggregation is effective.

[1] Our Python implementation is available in https://github.com/crowd4u/ci25-nishio

- Clarified the conditions under which using rationales in label aggregation is effective.

2 Related Work

2.1 Aggregation Methods Using Only Worker Responses

The Dawid-Skene (DS) method [4] is known as an effective aggregation method for crowdsourcing responses, and several extensions of the DS method have been proposed. One such extension is the OneCoin model, which estimates worker ability based only on accuracy instead of using the confusion matrix. The OneCoin model is more robust than the DS method and has a better convergence rate for the EM algorithm [32].

The GLAD method proposed by Whitehill et al. [31] estimates both worker ability and task difficulty, and it is particularly strong in situations involving adversarial workers or noise. Zhou et al. [33] assumed that the labels assigned by workers are generated based on a probability distribution over worker-task-label triplets and minimized the entropy of these probability distributions to estimate the true labels with better accuracy than the DS method.

The DS method performs estimation using the EM algorithm. However, there are proposals [14] to improve accuracy by implementing the DS method and its extended versions using Bayesian inference instead of the EM algorithm. In the classifier combination method proposed by Kim et al. [14], human labelers are treated as classifiers, and human-based label aggregation is performed by incorporating humans into the combination of classifiers.

However, these methods aggregate results based only on worker responses and do not consider rationales which workers provide to their answers.

2.2 Quality Control Using Worker Rationales

In research that uses worker rationales to improve crowdsourcing quality management, Inel et al. [10] reported that by having workers highlight the part of the text that serves as the basis for their judgment in a task determining whether a document and topic are related, the workers' accuracy improved. McDonnell et al. [19] proposed a method that uses the degree of agreement between the input rationales within a task to filter responses, where only responses with high agreement are used for aggregation, thus improving aggregation quality. Our method can be easily combined with the filtering approach, but since the filtering approach reduces the number of responses, it is not clear whether the combination will yield better results. McDonnell et al. [19] also discussed task designs that require workers to provide rationales for their responses, suggesting that a two-stage task design, in which one worker inputs responses and rationales while another evaluates others' inputs, is effective. Since the research question in this study is independent of task design, task design is considered outside the scope of this work.

Kida et al. [13] proposed a method for deriving formal logic proofs from the rationales input by workers. Formal logic refers to a mathematical framework for proof construction. Although the method produces the aggregated result as a side effect of generating the proof, constructing a proof requires a complex workflow involving a variety of tasks and a large number of worker contributions. This is a completely different approach to our worker result aggregation.

These studies utilize worker rationales for crowdsourcing quality management, but none of them propose aggregation methods that directly use the rationales.

2.3 Aggregation Methods Utilizing Additional Information

As an aggregation method using additional information, Oyama et al. [22] revealed that while there is a positive correlation between the average confidence level of workers' responses and the workers' accuracy, there are many workers whose confidence level deviates from their actual accuracy, such as overconfident or underconfident workers. They proposed an aggregation method that uses the workers' reported confidence levels in their labels. These studies are related to this research in that they also involve aggregation methods using additional information, but our proposed aggregation method differs by using workers' rationales in addition to their responses as the input information.

Additionally, Goyal et al. [8] proposed an aggregation method that uses behavioral data of workers as additional information. Here, behavioral data refers to the actions workers perform during a task, such as clicks, mouse movements, or page scrolling.

3 Preliminaries and Problem Setting

Table 1 shows the notation used in this paper. Using these notations, we first explain the Dawid-Skene method, which forms the foundational knowledge for this study, and then describe the problem setting addressed in this paper.

3.1 Aggregation in Crowdsourcing

Quality control is a major challenge in crowdsourcing [3]. Since workers come from a large, unspecified group, their abilities are inherently unstable, and high-quality responses are not always guaranteed. Furthermore, the response process of workers is often opaque, leading to the presence of workers who submit careless answers for easy rewards and spam workers who provide random responses to earn compensation.

There are several approaches to quality control such as task design and aggregation methods. This study focuses on aggregation methods. Task design involves improving quality control by carefully designing task instructions, descriptions, and the user interface of the task screen. In contrast, aggregation methods assign multiple workers to the same task and consolidate their responses into a single aggregated result.

The simplest aggregation method is majority voting, where the label with the highest number of responses is adopted as the final result. However, due to the variability in worker ability and motivation, majority voting does not always yield high accuracy. In particular, when there is a skewed distribution of worker ability—where a small number of high-ability workers are outnumbered by a large number of low-ability workers—majority voting may lead to a decline in aggregation quality. To address this issue, methods have been proposed that estimate not only the aggregated result but also the individual abilities of workers to improve the overall aggregation process.

3.2 Dawid-Skene Method

The Dawid-Skene (DS) method is an aggregation technique for multi-class classification, which predicts the true class of each task by using the EM algorithm while estimating the workers' abilities and the class's marginal distribution, represented by the confusion matrix. The graphical model of the DS method is shown in Figure 2.

The notations used in this section are shown in Table 1. Let the set of tasks be denoted by $\mathbb{I}$, the set of workers by $\mathbb{K}$, and the set of classes which workers can choose by $\mathbb{J}$. For example, in a task where images are classified into $\{Octopus, Squid, Jellyfish\}$, we have $\mathbb{J} = \{Octopus, Squid, Jellyfish\}$. The label assigned by worker k to task i is denoted as $y_i^{(k)} \in \mathbb{J}$. The true class of task i is represented by $t_i \in \mathbb{J}$. The expected value of the true class of a task i is j is denoted by E_{ij}. The number of times worker k answers class j for task i is denoted by $n_{ij}^{(k)}$. In our setting, $n_{ij}^{(k)}$ is 0 or 1 since we assume the workers will not answer multiple times. The marginal probability of class j is denoted by $\rho_j = p(t_i = j)$, and the probability that worker k responds with class $l \in \mathbb{J}$ when the true class of a task is j is denoted by $\pi_{jl}^{(k)} = p(y_i^{(k)} = l \mid t_i = j)$.

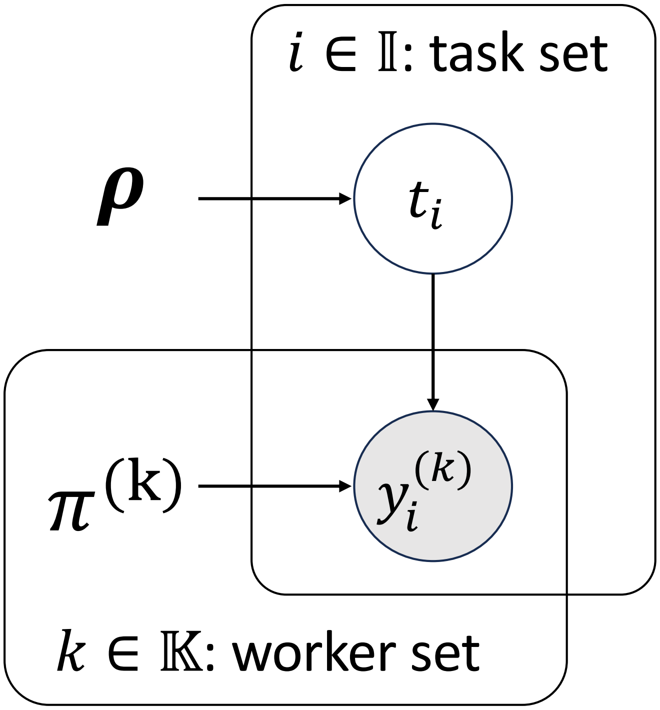


Fig. 2. The graphical model of the DS method

The objective of the DS method is to estimate the expected true label E_{ij} given the observed worker responses $n_{ij}^{(k)}$. To achieve this, the method estimates the latent variables: the class marginal probabilities ρ_j and the confusion matrix $\boldsymbol{\pi}^{(k)}$ for each worker k. The confusion matrix $\boldsymbol{\pi}^{(k)}$ is a $|\mathbb{J}| \times |\mathbb{J}|$ matrix where each element $\pi_{jl}^{(k)}$ represents the probability that worker k assigns class l when the true class is j. In the DS method, this confusion matrix is interpreted as a measure of worker ability.

The DS method is based on the EM algorithm, iteratively updating estimates through the following steps:

- E-step: Compute the expected value of the true label E_{ij}
- M-step: Estimate the latent variables, i.e., the class marginal probabilities ρ_j and the confusion matrices $\boldsymbol{\pi}^{(k)}$, using maximum likelihood estimation.

The computation for E_{ij} in the E-step is provided below.

$$E_{ij} = \frac{\rho_j \prod_{l \in \mathbb{J}} \prod_{k \in \mathbb{K}} (\pi_{jl}^{(k)})^{n_{il}^{(k)}}}{\prod_{q \in \mathbb{J}} \rho_q \prod_{l \in \mathbb{J}} \prod_{k \in \mathbb{K}} (\pi_{ql}^{(k)})^{n_{il}^{(k)}}} \tag{1}$$

The estimation of ρ_j and $\boldsymbol{\pi}^{(k)}$ in the M-step is also detailed below, where the latent variables are estimated using the maximum likelihood method.

$$\rho_j = \frac{\sum_{i \in \mathbb{I}} E_{ij}}{|\mathbb{I}|} \tag{2}$$

$$\pi_{jl}^{(k)} = \frac{\sum_{i\in\mathbb{I}} E_{ij} n_{il}^{(k)}}{\sum_{s\in\mathbb{J}} \sum_{i\in\mathbb{I}} E_{ij} n_{is}^{(k)}} \tag{3}$$

Since the DS method requires an initial value for E_{ij}, empirical studies suggest effective initialization strategies [4], which are introduced in the following section.

$$E_{ij} = \frac{\sum_{k\in\mathbb{K}} n_{ij}^{(k)}}{\sum_{l\in\mathbb{J}} \sum_{k\in\mathbb{K}} n_{il}^{(k)}} \tag{4}$$

3.3 Problem Setting

Table 1. Notations

Symbols	Description
$\mathbb{I}$	Set of tasks
$\mathbb{J}$	Set of classes
$\mathbb{K}$	Set of workers
$\mathbb{R}$	Set of rationales
$y_i^{(k)}$	The label assigned by worker k to task i
$e_i^{(k)}$	The rationale provided by worker k for task i.
t_i	True label of task i
E_{ij}	The expected value of t_i
$\pi_{jl}^{(k)}$	The probability that worker k labels a task as class l when its true class is j
$n_{ij}^{(k)}$	The number of times worker k assigned class j to task i
m_{ijr}	The number of times class j along with rationale r has been provided for task i

In this paper, we consider the problem where a worker $k \in \mathbb{K}$ assigns a label $j \in \mathbb{J}$ to a task $i \in \mathbb{I}$ and provides a rationale $r \in \mathbb{R}$ for their labeling decision. Table 1 presents the notations used throughout this paper. Let $\mathbb{I}$ denote the set of tasks, $\mathbb{J}$ the set of classification categories, $\mathbb{K}$ the set of workers, and $\mathbb{R}$ the set of rationales. Here, we treat rationales as categorical variables.

The label assigned by worker k to task i is denoted as $y_i^{(k)} \in \mathbb{J}$. The rationale provided by worker k for a task i is denoted as $e_i^{(k)} \in \mathbb{R}$. Let $n_{ij}^{(k)}$ denote the number of times worker k has assigned class j to task i, and let m_{ijr} represent the number of times class j along with rationale r has been provided for task i.

Our objective is to estimate the set of true labels $\boldsymbol{t} = \{t_i\}_{i\in\mathbb{I}}$ given the set of assigned labels $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i\in\mathbb{I}} = \{y_i^{(k)}\}_{i\in\mathbb{I},k\in\mathbb{K}}$ and the set of provided rationales $\boldsymbol{e} = \{\boldsymbol{e}_i\}_{i\in\mathbb{I}} = \{e_i^{(k)}\}_{i\in\mathbb{I},k\in\mathbb{K}}$.

4 Rationale-aware Aggregation

4.1 Model

This section introduces a probabilistic generative model that estimates the true label based on both the workers' responses and their rationales, which is the basis of our RAAG method.

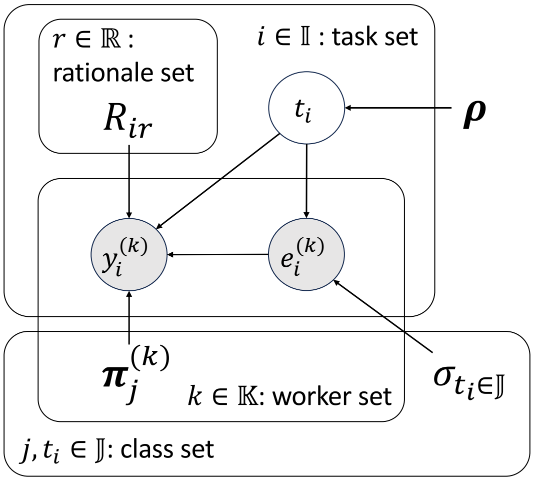

$r \in \mathbb{R}$:
rationale set
R_{ir}
$i \in \mathbb{I}$: task set
t_i
$\boldsymbol{\rho}$
$y_i^{(k)}$
$e_i^{(k)}$
$\boldsymbol{\pi}_j^{(k)}$
$k \in \mathbb{K}$: worker set
$\sigma_{t_i \in \mathbb{J}}$
$j, t_i \in \mathbb{J}$: class set

Fig. 3. The graphical model of RAAG

This model is characterized by the following joint distribution:

$$\begin{aligned} &p(\boldsymbol{t}, \boldsymbol{e}, \boldsymbol{y}) && (5)\\ &= \prod_{i \in \mathbb{I}} \prod_{k \in \mathbb{K}} p(y_i^{(k)} | \boldsymbol{e}_i, t_i) p(e_i^{(k)} | t_i) p(t_i) \end{aligned}$$

The graphical model of RAAG is shown in Figure 3.

First, the true label t_i of task i is generated according to the following categorical distribution with parameter $\boldsymbol{\rho}_j$. Here, similar to the Dawid-Skene model, we assume that the prior distribution of the true label is shared across all data.

$$p(t_i) = Cat(t_i | \boldsymbol{\rho}) \tag{6}$$

$$= \prod_{j \in \mathbb{J}} \rho_j^{\mathbb{1}[t_i = j]} \tag{7}$$

We assume that the rationale $e_i^{(k)}$ provided by worker k for task i is conditionally independent given the true label t_i. When the true label of task i is j, the rationale $e_i^{(k)}$ provided by worker k is generated according to the following categorical distribution with parameter $\boldsymbol{\sigma}_j$:

$$p(e_i^{(k)} | t_i) = Cat(e_i^{(k)} | \boldsymbol{\sigma}_{t_i}) \tag{8}$$

$$= \prod_{r \in \mathbb{R}} \sigma_{t_i, r}^{\mathbb{1}[e_i^{(k)} = r]} \tag{9}$$

The response $y_i^{(k)}$ from worker k for task i is generated according to a categorical distribution with parameter $\boldsymbol{\tau}_i^{(k)}$.

$$p(y_i^{(k)}|\boldsymbol{e}_i, t_i) = Cat(y_i^{(k)}|\boldsymbol{\tau}_i) \tag{10}$$

$$= \prod_{j\in\mathbb{J}} (\tau_{ij}^{(k)})^{\mathbb{1}[y_i^{(k)}=j]} \tag{11}$$

$\boldsymbol{\tau}_i^{(k)}$ is a vector of dimension $|\mathbb{J}|$ with elements $\tau_{ij}^{(k)}$. The value of $\tau_{ij}^{(k)}$ is calcuted using the worker's confusion matrix $\pi_{jl}^{(k)}$ and the strength of the rationale R_{ir} for task i by the following equation 12. $z_i^{(k)}$ is a normalization term.

$$\tau_{ij}^{(k)} = \frac{1}{z_i^{(k)}} \prod_{l\in\mathbb{J}} (\pi_{lj}^{(k)})^{\mathbb{1}[t_i=l]} \times \{1 - \prod_{r\in\mathbb{R}} (1 - R_{ir})^{m_{ijr}}\} \tag{12}$$

4.2 Parameter Estimation

We estimate the true label t_i from the observed labels $y_i^{(k)}$ and the provided rationales $e_i^{(k)}$.

The parameters of the model and the true labels are alternately estimated using the Expectation-Maximization (EM) algorithm.

In the E-step, we estimate the expected value E_{ij}. The initial value of E_{ij} is set to be the same as in the DS method.

$$E_{ij} = p(t_i = j \mid \boldsymbol{y}_i, \boldsymbol{e}_i) \tag{13}$$

Since $y_i^{(k)}$ and $e_i^{(k)}$ are observed, equation 14 holds.

$$p(t_i|\boldsymbol{y}_i, \boldsymbol{e}_i) \propto \prod_{k\in\mathbb{K}} p(y_i^{(k)}|\boldsymbol{e}, t_i) p(e_i^{(k)}|t_i) p(t_i) \tag{14}$$

Therefore, the expected value E_{ij} is estimated by the following equation 15. Z_i is a normalization term.

$$E_{ij} = \frac{1}{Z_i} \rho_j \{1 - \prod_{r\in\mathbb{R}} ((1 - \sigma_{jr})^{m_{ijr}})\} \prod_{k\in\mathbb{K}} \tau_{ij}^{(k)} \tag{15}$$

In the M-step, the parameters $\boldsymbol{\rho}$, $\{\boldsymbol{\sigma}_j\}_{j\in\mathcal{J}}$, $\{R_{ir}\}_{i\in\mathbb{I}, r\in\mathbb{R}}$ are estimated.

$\boldsymbol{\rho}$ is a vector of dimension $|\mathbb{J}|$ with elements ρ_j. The element ρ_j is estimated by the following equation 16.

$$\rho_j = \frac{\sum_{i\in\mathbb{I}} E_{ij}}{\sum_{j'\in\mathbb{J}} \sum_{i\in\mathbb{I}} E_{ij'}} \tag{16}$$

$\boldsymbol{\sigma}_j$ is a vector of dimension $|\mathbb{R}|$ with elements σ_{jr}. The value of σ_{jr} is estimated by the following equation 17.

$$\sigma_{jr} = \frac{\sum_{i\in\mathbb{I}} m_{ijr}}{\sum_{r'\in\mathbb{R}} \sum_{i\in\mathbb{I}} m_{ijr'}} \tag{17}$$

Here, the worker's confusion matrix $\pi_{jl}^{(k)}$ is estimated using equation 3 in the same manner as in the DS method. Furthermore, the strength of the rationale R_{ir} is estimated using equation 18.

$$R_{ir} = \frac{\sum_{j\in\mathbb{J}} m_{ijr} E_{ij}}{\sum_{r'\in\mathbb{R}} \sum_{j\in\mathbb{J}} m_{ijr'} E_{ij}} \tag{18}$$

Table 2. Rationale Groups in Experiment with Synthetic Data

	Group1			Group2			Group3
True label	A	B	C	D	E	F	G
0	0.8	0.1	0.1	0.4	0.4	0.2	1
1	0.1	0.8	0.1	0.2	0.4	0.4	1
2	0.1	0.1	0.8	0.4	0.2	0.4	1

5 Experiments with Synthetic Data

We conducted experiments with synthetic and real-world data sets to address RQ1 and RQ2. In particular, to verify the conditions under which using rationales in label aggregation is effective, we prepared multiple conditions for the synthetic data and compared the aggregation quality for each condition.

5.1 Data Generation

We constructed a data model for synthetic data following the findings in the paper by McDonnell et al. [19]: *highly skilled workers provide the same rationale*. Based on this findings, we consider that rationales with higher agreement among workers in each task are *stronger* rationales, and we generate multiple synthetic data sets with rationales with different strengths.

Dataset Structure. Each synthetic dataset consists of two components; the first component is a set of $|\mathbb{I}| = 2520$ 3-class classification tasks, where for each $i \in \mathbb{I}$, its answer is one of the classes $\mathbb{J} = \{0, 1, 2\}$, with 840 tasks for each class. The second component is the set of worker responses to the $|\mathcal{I}|$ tasks. The number of worker responses varies according to data generation parameters. For example, if the redundancy is 3, we have 7560 worker responses to the 2520 tasks. Each worker response is connected to one worker id. The number of workers is detenmined by the number of tasks assigned per worker parameter. If we want to assign 10 tasks to each worker, the dataset assumes 7560/10=756 workers. We use $\mathbb{K}$ to denote the number of workers.

Generating Worker Responses. Each worker response contains the answer for a task and its rationale. The rationales provided by worker responses are classified into 7 types (denoted by $\mathbb{R} = \{A, B, C, D, E, F, G\}$) and these 7 types of rationales are grouped into 3 groups $\mathbb{R}_1 \oplus \mathbb{R}_2 \oplus \mathbb{R}_3 = \{A, B, C\} \oplus \{D, E, F\} \oplus \{G\}$.

Each worker $k \in \mathbb{K}$ is affiliated with Group g and the worker selects a rationale $r \in \mathbb{R}_g$. The worker probabilistically determines the rationale to select based on their affiliated group and the true label of the task. The true accuracy of worker k for their true label is denoted as $\alpha^{(k)}$.

The confusion matrix $\boldsymbol{\gamma}^{(k)}$ for worker k is shown in Table 3, where $\gamma_{jl}^{(k)} = p(y_i^{(k)} = l|t_i = j)$ represents the probability that worker k answers l when the true label of the task is j.

The selection probabilities $\beta_{jr}^{(g)} = p(e_i^{(g)} = r|t_i = j)$ of rationale r for the true label j in group g are shown in Tables 2. For example, a worker in group 1, when the true label is 1, will input rationale A with a probability of 0.1, rationale B with a probability of 0.8, and rationale C with a probability of 0.1.

Therefore, each group can be characterized as follows: Workers in group 1 are more likely to provide rationales that match the true label. Workers in group 2 are less likely to match the rationale compared to workers in group A. In other words, workers in group 1 are skilled and provide strong rationales, while workers in group 2 are less skilled and provide weaker rationales. Additionally, workers in group 3 only provide rationale G, regardless of the true label. This is intended to represent spam workers who provide meaningless information without considering the task content.

Table 3. Confusion Matrix in Experiment with Synthetic Data

	0	1	2
0	$\alpha^{(k)}$	$\frac{1-\alpha^{(k)}}{2}$	$\frac{1-\alpha^{(k)}}{2}$
1	$\frac{1-\alpha^{(k)}}{2}$	$\alpha^{(k)}$	$\frac{1-\alpha^{(k)}}{2}$
2	$\frac{1-\alpha^{(k)}}{2}$	$\frac{1-\alpha^{(k)}}{2}$	$\alpha^{(k)}$

The response $y_i^{(kg)} \in \mathbb{J}$ of each synthetic worker k in group g is determined probabilistically based on the true label t_i, the worker's accuracy $\alpha^{(k)}$, and the selected rationale r. The probability that the worker's response is $y_i^{(kg)} = l$ is given by the equation (19).

$$\begin{aligned} p(y_i^{(kg)} = l \mid t_i = j) &= \frac{p(y_i^{(k)} = l, e_i^{(g)} = r \mid t_i = j)}{p(e_i^{(g)} = r | t_i = j)} \\ &= \frac{p(e_i^{(g)} = r \mid t_i = j) p(y_i^{(k)} = l \mid t_i = j)}{\sum_{l' \in \mathbb{J}} p(e_i^{(g)} = r | t_i = j) p(y_i^{(k)} = l' \mid t_i = j)} \\ &= \frac{\beta_{jr}^{(g)} \gamma_{jl}^{(k)}}{\sum_{l \in \mathbb{J}} \beta_{jr}^{(g)} \gamma_{jl}^{(k)}} \end{aligned} \tag{19}$$

5.2 Generated Synthetic Datasets

To address RQ2, we compared RAAG and DS in terms of four perspectives: redundancy, the number of tasks assigned per worker, average accuracy, and group ratio. For a comparison from each perspective, we generated a baseline dataset and other datasets each of which is generated with the same parameter sets as the baseline except parameters that control the perspective.

- **Redundancy**: We explored the effect of changing the number of workers who respond to a single task. If the redundancy is 3, it means that three workers will respond to one task. In the experiment, 5 will be used as the baseline, and from 2 to 10 will be used as comparison conditions.
- **Number of tasks assigned per worker**: The effect of the number of tasks each worker responds to. If the number is 5, each worker responds to 5 tasks. In the experiment, 5 will be used as the baseline, and from 2 to 10 will be used as comparison conditions.
- **Average accuracy**: The effect of the average accuracy of task results. In the experiment, 0.6 will be used as the baseline average accuracy, and from 0.4 to 0.8 in increments of 0.1 will be used as comparison conditions. To implement the difference in accuracy among workers, each worker's accuracy was set to the average accuracy plus a random value of ±0.2.
- **Group ratio**: The effect of the ratio of the number of workers belonging to each group g. For example, when the group ratio is 1:1:2, and there are 400 workers, 100 workers belong to group 1 and group 2, and 200 workers belong to group 3. In the experiment, 1:1:1 will be used as the baseline, and 1:0:0, 0:1:0, 0:0:1, 2:1:1, 1:2:1, and 1:1:2 will be used as comparison conditions.

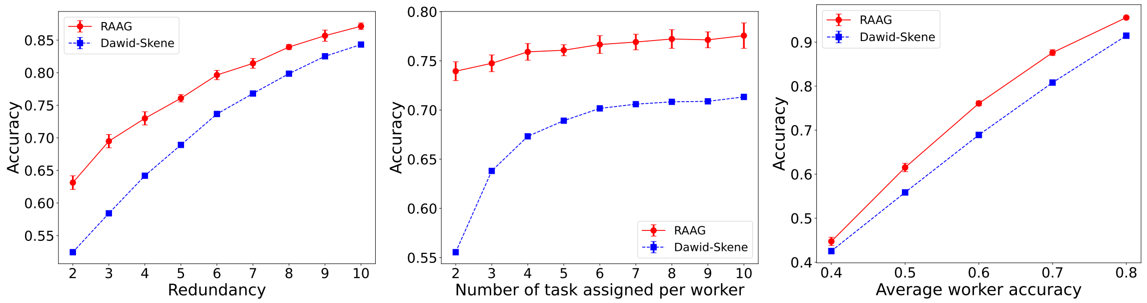


Fig. 4. Accuracy under each condition in the experiment with synthetic data

Table 4. Accuracy with Varying Group Ratio

Group Ratio	DS	RAAG
1 : 1 : 1 (baseline condition)	0.721	0.782
1 : 0 : 0	0.721	0.797
0 : 1 : 0	0.721	0.774
0 : 0 : 1	0.721	0.727
2 : 1 : 1	0.721	0.792
1 : 2 : 1	0.721	0.783
1 : 1 : 2	0.721	0.771

By setting the group ratio to 1:0:0, 0:1:0, or 0:0:1, synthetic data consisting only of each group can be created. By examining the aggregation quality for each group, we can verify the effectiveness of RAAG based on the nature of strong or weak rationales.

Additionally, by setting the group ratio to 2:1:1, 1:2:1, or 1:1:2, synthetic data can be created that represents a situation where one group has a higher proportion while the groups are mixed. This synthetic data allows us to investigate how differences in the proportions of strong or weak rationales in a mixed group affect RAAG.

For the experiment, we generated 10 data sets for each set of parameters, because the generation process is not deterministic. The experiment results show the average of the aggregation results on the average of the ten data sets for each condition (i.e., each set of parameters).

5.3 Results of the Synthetic Data Experiment

Figure4 and Tables4 show the results. All parameters not explicitly shown are the same as those in the baseline condition. The results for the baseline using synthetic data are shown in the first row of Table 4. In the baseline setting, RAAG improved the accuracy by 0.061 compared to the DS method.

The experimental results with varying group ratios are shown in Table 4. When the group ratio was set to 1:0:0, 0:1:0, and 0:0:1, RAAG improved the accuracy by 7.6%, 5.3%, and 0.6% compared to the DS method, respectively. When the group ratio was set to 2:1:1, 1:2:1, and 1:1:2, RAAG improved the accuracy by 7.1%, 6.2%, and 5.0% compared to the DS method, respectively.

5.4 Discussion

5.4.1 Redundancy. Overall, as the redundancy level increases, the difference between the RAAG and DS methods decreases. In general, setting a high redundancy level tends to improve the accuracy of aggregation regardless of the aggregation method, whereas setting a low redundancy level tends to decrease accuracy. Although both the DS and RAAG exhibit lower accuracy at low redundancy levels, RAAG achieves higher accuracy than DS and mitigates the negative impact of reduced redundancy. This suggests that RAAG, by utilizing supporting information, can yield more accurate aggregated results than DS even in situations where the correct workers are in the minority. Furthermore, at high redundancy levels, DS can achieve high accuracy even without considering supporting information through RAAG, which likely explains the minimal difference between RAAG and DS in such cases.

5.4.2 Number of Task Assignments. In the DS method, reducing the number of tasks assigned per worker led to a decrease in accuracy. However, the RAAG was less affected by a smaller number of assignments compared to the DS method. It has been pointed out that models that estimate worker ability individually, such as the DS method, may struggle to accurately estimate worker ability when the number of assignments is low, resulting in a decline in accuracy [9, 28]. Unlike the DS method, the RAAG estimates the true label not only based on worker ability but also by incorporating rationales. Consequently, it was able to maintain relatively high accuracy even in situations where worker ability could not be accurately estimated.

5.4.3 Average Accuracy. The difference in accuracy between the DS method and RAAG is the largest when the average accuracy is 0.6. Overall, the difference decreases as the average accuracy deviates from 0.6. When the average accuracy is set to a low value of 0.4, the improvement in accuracy achieved by RAAG is relatively small. This is likely influenced by the method used in the proposed approach to estimate the strength of rationales. In the RAAG, the strength of rationales is calculated based on the expected values for each class in a given task. Since this method adopts the majority vote-based expected values as the initial values, it is presumed that under conditions of low average accuracy, the initial expected value of the true label is also low, leading to inaccurate estimation of the strength of rationaless.

5.4.4 Group Ratio. When the group ratio was 1:0:0 (all workers are in group 1), all workers input strong rationales, and RAAG's estimation of the true label using the rationales worked well, leading to a higher accuracy improvement compared to the baseline.

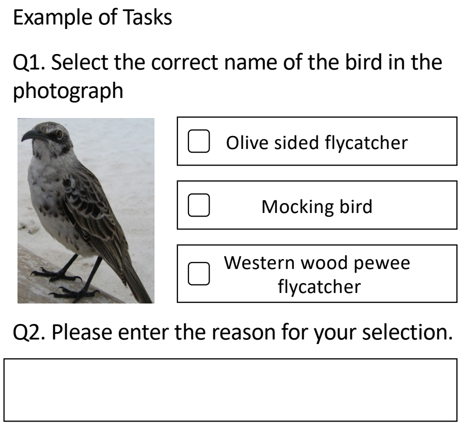
When the group ratio was 0:1:0 (all workers are in group 2), the improvement in accuracy was lower than the baseline. In the baseline, all groups are mixed, and aggregation is performed based on strong rationales while estimating the strength of the rationales. However, when only workers from group 2 are present, only weak rationales are used for aggregation, leading to a lower improvement in accuracy compared to the baseline.

When the group ratio was 0:0:1 (all workers are in group 3), the accuracies of RAAG and the DS method were almost identical. This is because all workers input the same rationales, so there is no meaningful aggregation of rationales, resulting in no difference between the two methods.

When the group ratio was 2:1:1, the accuracy improvement was slightly higher compared to the baseline. This is likely due to the increased proportion of strong rationales, which influenced the result. Additionally, when the group ratio was 1:2:1, the accuracy improvement was almost the same as the baseline. This suggests that even in situations where both strong and weak rationales are mixed, RAAG effectively estimates the strength of the rationales and performs aggregation based on the strong rationales.

Table 5. Classification Classes and Datasets

Target Category	Classification Classes	Dataset
Birds	Olive sided flycatcher, Mocking bird, Western wood pewee Flycatcher	Caltech-UCSD Birds-200-2011 [29]
Dog Breeds	Flat-coated Retriever, Curly-coated Retriever, Newfoundland	Stanford Dogs Dataset [12]
Plants	Black-grass, Loose-silky-bent, Common-wheat	V2 Plant Seeding Dataset [7]


Example of Tasks

Q1. Select the correct name of the bird in the photograph

□ Olive sided flycatcher

□ Mocking bird

□ Western wood pewee flycatcher

Q2. Please enter the reason for your selection.

Fig. 5. Task design

When the group ratio was 1:1:2, the accuracy improvement was lower than the baseline. However, even in situations where the proportion of workers from group 3, who provide meaningless rationales, was higher, RAAG still achieved a higher accuracy compared to the DS method. This suggests that RAAG is robust against spam workers who do not provide meaningful rationales.

6 Experiments with Real-World Datasets

This section addresses RQ1 using real-world datasets. In the experiment, three sets of image classification tasks are performed via crowdsourcing: bird image classification tasks, dog breed image classification tasks, and plant image classification tasks.

6.1 Datasets

The datasets and task categories used in this experiment are shown in Table 5. The bird classification task uses the `Caltech-UCSD Birds-200-2011` [29], the dog breed classification task uses the `Stanford Dogs Dataset` [12], and the plant classification task uses the `V2 Plant Seeding Dataset` [7].

6.2 Experimental Setup

For each task, the target classes were limited to three, and 50 images were randomly selected from the dataset for each class and used in the experiment. The redundancy and the number of task assignments to workers were set to 10. To investigate the effect of different redundancy on RAAG, data were randomly sampled to match the specified redundancy, and the aggregation was performed using both RAAG and the existing DS method.

6.3 Task Design

The Yahoo! Crowdsourcing platform[2] was used as the crowdsourcing platform for labeling tasks. The total number of tasks for labeling via crowdsourcing is

$$\begin{aligned} &3(\text{classification tasks}) \times 50(\text{images per class}) \times 3(\text{classes}) \\ &= 450 \text{ tasks} \end{aligned}$$

with redundancy set to 10 and the number of assignments set to 10. Therefore, the number of workers required to complete the tasks is $450(\text{tasks}) \times 10(\text{redundancy})/10(\text{assignments}) = 450$ workers.

Workers eligible to perform the tasks were limited to those aged 18 and above. Workers who completed the tasks were rewarded with the equivalent of 65 JPY via the platform. No qualification question, where tasks would not be considered complete unless the correct answer was provided, was utilized in this experiment.

Obtaining rationales. Donahue et al. [5] proposed two methods for inputting rationales in image labeling tasks: (1) selecting the region of the image that the worker considers most important for labeling using a bounding box, and (2) selecting the attributes of the image that the worker considers most important when labeling. In the field of image recognition addressed by Donahue et al., labeling by adding bounding boxes to the image is effective. However, in order to address a more general problem setting, our RAAG does not use the first approach. Moreover, the second approach assumes that attribute information is available in the dataset, but in this study, we aim to collect rationales in a way that is independent of the dataset. Therefore, in this experiment, workers are asked for the rationales for their answers in a free-text field, which is then preprocessed and converted into categorical variables for use in aggregation as we will explain next.

Figure5 illustrates the task design based on the above. It asks the worker to choose one of the classes and fill in a free-text field for the rationale for the choice.

6.4 Preprocessing of Collected rationales

The rationales collected through free text input are converted into categorical variables using the following steps:

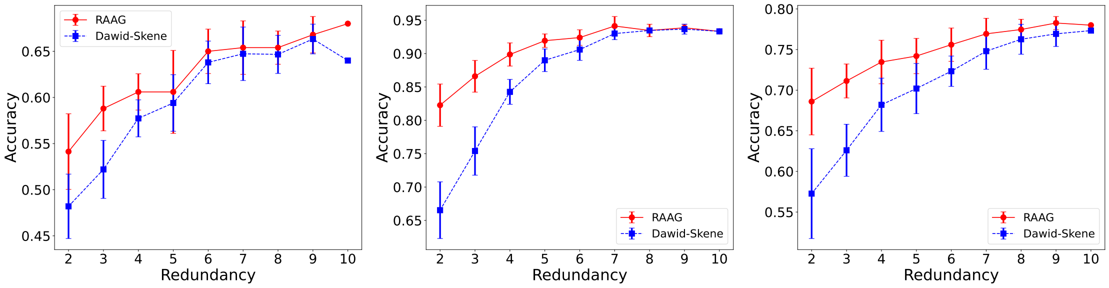
(1) Convert each piece of rationales into a vector. For this conversion, we use the Python library `sentence-transformers 3.3.1` [24] and the multilingual model `sts-b-xlm-r-multilingual` [25]. The resulting vector is a 768-dimensional dense vector.

(2) Cluster the vectors obtained in (1) and consider the data belonging to each cluster as the same ratinales. We use X-means [23] based on Bayesian Information Criterion for clustering.

6.5 Results & Discussion

The experimental results are shown in Figures 6a, 6b, and 6c. The error bars represent the standard deviation.

[2] https://crowdsourcing.yahoo.co.jp/



(a) Birds

(b) Dogs

(c) Plants

Fig. 6. Accuracy of Image Classification Task

In all tasks, when redundancy is low, the RAAG remarkably improved the accuracy compared to DS;

it achieved accuracy improvements of up to 6.6%, 15.7%, and 11.3% on the Birds, Dogs, and Plants datasets, respectively, all under low-redundancy conditions (i.e., redundancy levels of 2 or 3). This result supports the results of synthetic data experiments where the RAAG is effective in low redundancy conditions. Furthermore, as redundancy increases, the difference between the RAAG and DS becomes smaller, which also supports the results of synthetic data experiments where the difference decreases with higher redundancy.

In the real-world data experiments, similar to the synthetic data experiments, the RAAG improved the accuracy. This suggests that label aggregation using the workers' response rationales can enhance the quality of the aggregation results.

While RAAG holds promise for applications across a wide range of domains, challenges remain in how supporting evidence is handled. RAAG treats supporting evidence as categorical variables and there is considerable room for improvement in this aspect. This is because approaches like vectorization and clustering of natural language we used for our implementation can lose subtle and nuanced information contained in complex textual evidence. In the experiment, we used SBERT just because our data set contains texts in multiple languages and it is a popular tool that can deal with them, not because it is the best tool. Using better tools might further improve the performane of RAAG. As future work, exploring more sophisticated methods for representing and integrating natural language evidence, such as leveraging large language models for semantic interpretation or utilizing contextual embeddings that better preserve linguistic nuance, may lead to more faithful and effective aggregation, enhancing both the performance and transparency of RAAG.

7 Conclusion

This paper investigated the effects of exploiting workers' response rationales in label aggregation in crowdsourcing. For this purpose, we validated the effectiveness of RAAG, which is a DS-based method that estimates the true labels as well as the workers' abilities and the strength of the rationales based on the EM algorithm, through experiments using synthetic data and real-world data. Incorporating supporting evidence into the aggregation process makes it a promising approach for applications in domains where high reliability and accountability are critical.

The experimental results with synthetic and real-world data sets showed that RAAG is particularly effective under challenging conditions, such as low redundancy, limited responses per worker, and the presence of spammers. In particular, we found that RAAG performs label aggregation with higher accuracy up to 15.7% compared to DS, under the condition with the lowest redundancy.

Acknowledgments

A part of this research is supported by JSPS Grants-in-Aid for Scientific Research (JP22H00508, JP23K28095, JP23K21723, JP25K03415), and JST CREST (Grant Number JPMJCR22M2).

References

[1] Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Thanh Tran Duc. 2011. Repeatable and Reliable Search System Evaluation Using Crowdsourcing. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 923–932. doi:10.1145/2009916.2010039

[2] António Correia, Diogo Guimarães, Dennis Paulino, Shoaib Jameel, Daniel Schneider, Benjamim Fonseca, and Hugo Paredes. 2021. AuthCrowd: Author Name Disambiguation and Entity Matching Using Crowdsourcing. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 150–155. doi:10.1109/CSCWD49262.2021.9437769

[3] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1 (Jan. 2018), 7:1–7:40. doi:10.1145/3148148

[4] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 20–28. doi:10.2307/2346806 jstor:2346806

[5] Jeff Donahue and Kristen Grauman. 2011. Annotator Rationales for Visual Recognition. In *2011 International Conference on Computer Vision*. 1395–1402. doi:10.1109/ICCV.2011.6126394

[6] GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. 2023. Gamified Crowdsourcing for Idiom Corpora Construction. *Natural Language Engineering* 29, 4 (July 2023), 909–941. doi:10.1017/S1351324921000401

[7] Thomas Mosgaard Giselsson, Rasmus Nyholm Jørgensen, Peter Kryger Jensen, Mads Dyrmann, and Henrik Skov Midtiby. 2017. A Public Image Database for Benchmark of Plant Seedling Classification Algorithms. doi:10.48550/arXiv.1711.05458 arXiv:1711.05458 [cs]

[8] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6 (June 2018), 41–49. doi:10.1609/hcomp.v6i1.13331

[9] Hideaki Imamura, Issei Sato, and Masashi Sugiyama. 2018. Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Clustering Model. doi:10.48550/arXiv.1802.04551 arXiv:1802.04551 [stat]

[10] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying Topical Relevance with Evidence-based Crowdsourcing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 1253–1262. doi:10.1145/3269206.3271779

[11] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. 2011. Crowdsourcing for Book Search Evaluation: Impact of Hit Design on Comparative System Ranking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 205–214. doi:10.1145/2009916.2009947

[12] A. Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2012. Novel Dataset for Fine-Grained Image Categorization : Stanford Dogs.

[13] Kai Kida, Hiroyoshi Ito, Masaki Matsubara, Nobutaka Suzuki, and Atsuyuki Morishima. 2022. Aggregating Crowd Intelligence over Open Source Information: An Inference Rule Centric Approach. In *The Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP2022)*.

[14] Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian Classifier Combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. PMLR, 619–627.

[15] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. doi:10.1145/1357054.1357127

[16] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in Computer Vision. *Foundations and Trends® in Computer Graphics and Vision* 10, 2 (2016), 103–175. doi:10.1561/0600000073 arXiv:1611.02145 [cs]

[17] Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. 2021. Capturing Covertly Toxic Speech via Crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang (Eds.). Association for Computational Linguistics, Online, 14–20.

[18] Xiaotian Lu, Arseny Tolmachev, Tatsuya Yamamoto, Koh Takeuchi, Seiji Okajima, Tomoyoshi Takebayashi, Koji Maruhashi, and Hisashi Kashima. 2021. Crowdsourcing Evaluation of Saliency-Based XAI Methods. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano (Eds.). Springer International Publishing, Cham, 431–446. doi:10.1007/978-3-030-86517-7_27

[19] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4 (Sept. 2016), 139–148. doi:10.1609/hcomp.v4i1.13287

[20] Nina Mouhammad, Johannes Daxenberger, Benjamin Schiller, and Ivan Habernal. 2023. Crowdsourcing on Sensitive Data with Privacy-Preserving Text Rewriting. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, Jakob Prange and Annemarie Friedrich (Eds.). Association for Computational Linguistics, Toronto, Canada, 73–84. doi:10.18653/v1/2023.law-1.8

[21] Allard Oelen, Markus Stocker, and Sören Auer. 2024. Creating and Validating a Scholarly Knowledge Graph Using Natural Language Processing and Microtask Crowdsourcing. *International Journal on Digital Libraries* 25, 2 (June 2024), 273–285. doi:10.1007/s00799-023-00360-7

[22] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. 2013. Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, Beijing, China, 2554–2560.

[23] Dan Pelleg and Andrew W. Moore. 2000. X-Means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 727–734.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. doi:10.48550/arXiv.1908.10084 arXiv:1908.10084 [cs]

[25] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. doi:10.48550/arXiv.2004.09813 arXiv:2004.09813 [cs]

[26] Martin W.P Savelsbergh and Marlin W. Ulmer. 2022. Challenges and Opportunities in Crowdsourced Delivery Planning and Operations. *4OR* 20, 1 (March 2022), 1–21. doi:10.1007/s10288-021-00500-2

[27] Ashutosh Timilsina and Simone Silvestri. 2023. E-Uber: A Crowdsourcing Platform for Electric Vehicle-based Ride- and Energy-sharing. In *2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. 359–365. doi:10.1109/MASS58611.2023.00051

[28] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-Based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. Association for Computing Machinery, New York, NY, USA, 155–164. doi:10.1145/2566486.2567989

[29] Catherine Wah, Branson Steve, Welinder Peter, Perona Pietro, and Belongle Serge. 2011. *Caltech-UCSD Birds-200-2011*. Technical Reports CNS-TR-2011-001. California Institute of Technology.

[30] Peter Washington, Haik Kalantarian, Jack Kent, Arman Husic, Aaron Kline, Emilie Leblanc, Cathy Hou, Cezmi Mutlu, Kaitlyn Dunlap, Yordan Penev, Nate Stockham, Brianna Chrisman, Kelley Paskov, Jae-Yoon Jung, Catalin Voss, Nick Haber, and Dennis P. Wall. 2021. Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. *Cognitive Computation* 13, 5 (Sept. 2021), 1363–1373. doi:10.1007/s12559-021-09936-4

[31] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, Vol. 22. Curran Associates, Inc.

[32] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. 2014. Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing. doi:10.48550/arXiv.1406.3824 arXiv:1406.3824 [stat]

[33] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 2195–2203.