Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work

MIKE SCHAEKERMANN, University of Waterloo, Canada
JOSLIN GOH, University of Waterloo, Canada
KATE LARSON, University of Waterloo, Canada
EDITH LAW, University of Waterloo, Canada

Crowdsourced classification of data typically assumes that objects can be unambiguously classified into categories. In practice, many classification tasks are ambiguous due to various forms of disagreement. Prior work shows that exchanging verbal justifications can significantly improve answer accuracy over aggregation techniques. In this work, we study how worker deliberation affects resolvability and accuracy using case studies with both an objective and a subjective task. Results show that case resolvability depends on various factors, including the level and reasons for the initial disagreement, as well as the amount and quality of deliberation activities. Our work reinforces the finding that deliberation can increase answer accuracy and the importance of verbal discussion in this process. We contribute a new public data set on worker deliberation for text classification tasks, and discuss considerations for the design of deliberation workflows for classification.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Human computer interaction (HCI)**; *Collaborative interaction*; *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: Inter-rater disagreement; Deliberation; Ambiguity

ACM Reference Format:
Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (November 2018), 19 pages. https://doi.org/10.1145/3274423

1 INTRODUCTION

Classification is a prevalent task in crowdsourcing as well as many real-world work practices. A common assumption for many classification tasks is that objects can be *unambiguously* classified into categories, and that the quality of the labeled data can be measured by the extent to which annotators agree with one another. As a result, most post-processing techniques designed to filter or aggregate labeled data interpret inter-rater disagreement as "noise in the signal" originating from human mistakes. In practice, many classification tasks are ambiguous, and disagreement can happen for various reasons including missing context, imprecise questions, contradictory evidence, and multiple interpretations arising from diverse levels or kinds of annotator expertise [14].

The independence of individual assessments has traditionally been considered a prerequisite for leveraging the 'wisdom of the crowd', but recent findings from crowdsourcing [7] and social

Authors' addresses: Mike Schaekermann, University of Waterloo, Waterloo, Canada, mschaeke@uwaterloo.ca; Joslin Goh, University of Waterloo, Waterloo, Canada, jtcgoh@uwaterloo.ca; Kate Larson, University of Waterloo, Waterloo, Canada, kate.larson@uwaterloo.ca; Edith Law, University of Waterloo, Waterloo, Canada, edith.law@uwaterloo.ca.

2573-0142/2018/11-ART154 $15.00
https://doi.org/10.1145/3274423

behavioural research [28] have found that deliberation can help improve answer quality. Drapeau et al. [7] showed that crowdsourcing workflows enabling workers to exchange justifications and to reconsider their assessments based on each other's arguments, can improve accuracy over output aggregation techniques. These prior works, however, have focused mostly on answer *correctness*, with the a priori assumption that each disagreement can be resolved to one correct answer.

In this work, we take the stance that inter-rater disagreement carries valuable information [2, 8, 18] and that deliberation should not always lead to a unanimous consensus. In particular, we investigate factors that contribute to the *resolvability* of a case. We conducted a study on two text classification tasks—a sarcasm classification task, which has been shown to be inherently ambiguous [9], and a semantic relation extraction task, where objective ground truth is available [7]—to investigate how deliberation affects resolvability. Our key contributions are:

(1) We study how the deliberation outcomes differ depending on task subjectivity.
(2) We present observations showing that the *resolvability* of an ambiguous case depends on the reasons for and level of initial disagreement, the amount and quality of the deliberation activities, as well as the task and case characteristics.
(3) We publish a new *dataset on worker deliberation* in text classification tasks, including all original and revised classifications as well as the deliberation dialogues.

The rest of this paper describes the related work, details the deliberation workflow designed and implemented for this study, outlines the experimental procedure and findings, and concludes with a discussion of applications and design considerations.

2 BACKGROUND

2.1 Sources of Inter-rater Disagreement

An early theoretical account identified three types of expert disagreement [27]—*personality-based* disagreement that arises due to incompetence, venality or ideology of the expert, *judgment-based disagreement* that arises due to missing information, or *structural disagreement* that arises due to experts adopting different problem definitions or organizing principles. Garbayo [12] distinguished *verbal* disagreement, i.e., discrepancy in terminology leading to misunderstanding, from *legitimate* disagreement that arises from experts having different interpretations of the same evidence.

Prior work has explored the issue of disagreement in crowd work. In visual question answering tasks, Gurari and Grauman [14] found that disagreement can be attributed to ambiguous and subjective questions, insufficient or ambiguous visual evidence, differing levels of annotator expertise, and vocabulary mismatch. Chang et al. [2] found that workers disagree because of incomplete or ambiguous category definitions, and proposed to shift efforts from a priori creation of label guidelines to post-hoc analysis of conceptual structures generated by the crowd. Kairam and Heer [18] introduced a clustering-based technique to identify subgroups of workers with diverging, but equally valid interpretations of the same (entity annotation) task. Their work shows that disagreement can depend on how conservatively or liberally workers interpret category definitions.

2.2 Rationale and Argumentation in Crowdsourcing

Prior work in crowdsourcing has explored ways to elicit explanations, or *rationale*, from annotators [7, 26]. For example, for search result relevance rating, McDonnell et al. [26] collected annotator rationale in the form of highlighted text in web pages, which was reviewed and fixed by a second worker. In a similar study on relation extraction by Drapeau et al. [7], workers were asked to asynchronously justify their assessments and reconsider them when confronted with counter-arguments from other workers. Results show that this form of asynchronous argumentation increased label quality compared to independent annotation under various budget constraints.

2.3 Protocols for Group Decision Making

A seminal protocol on structured decision making is the Delphi method [3], where experts provide, justify and reconsider their estimates through questionnaires in multiple rounds. A facilitator controls the information flow by summarizing estimates and filtering out irrelevant justification content at each round. The Delphi process ends after a fixed number of rounds or when unanimous consensus is reached. The key characteristics of the Delphi method are anonymity of the participants, avoidance of any direct interaction among group members, as well as structured and curated information flow as implemented by the facilitator. The Delphi method is typically used for forecasting, policy making and other types of complex decision making processes. Later versions of the Delphi method, e.g., by Hartman and Baldin [15], make use of computer-supported communication to facilitate remote collaboration, larger groups and asynchronous interaction. Like the Delphi method, our deliberation workflow ensures anonymity of the participants and imposes a round-based structure on the process of providing and reconsidering assessments. However, unlike Delphi, our workflow does not discard information between rounds, as we want workers to deliberate based on the full range of reasons for disagreement.

Group deliberation is a typical method for generating high-quality answers in expert domains such as medicine [13, 20, 32]; however, little work has shown under what circumstances group deliberation is resolvable or produces better decisions. Several works explored factors that influence the process and outcomes of group deliberation. Nemeth [29] found that when jurors are required to reach a unanimous decision, there is more conflict, more changes in assessments, and higher confidence in the final verdict reported by members of the group. Solomon [36] sees conflict as an important feature of any effective deliberation system. He argues that dissent is both necessary and useful—as "dissenting positions are associated with particular data or insights that would be otherwise lost in consensus formation"—and criticizes procedures that push deliberators to reach consensus. Instead, he advocates for a structured deliberation procedure that avoids the undesired effects of *groupthink* [17]—the tendency to agree with the group by suppressing dissent and appraisal of alternatives—by actively encouraging dissent, organizing independent subgroups to discuss the same problem, and ensuring diversity of group membership. Kiesler and Sproull [19] found that time limits imposed on deliberation tend to decrease the number of arguments exchanged and to polarize discussions. The authors suggest the use of voting techniques or explicit decision rules to structure the deliberation timeline. Our work draws inspiration from these prior works, by having multiple groups deliberate on the same case, by incentivizing workers to adamantly argue their point, and by structuring the decision process using a step-wise voting technique.

2.4 Online Deliberation Systems

Building on some of these early theoretical results, online deliberation systems have been developed and validated in various domains, including public deliberation [22, 23], on-demand fact checking [21], political debate [10] and knowledge base generation [41]. For example, ConsiderIt [10, 22] is a platform for supporting public deliberation on difficult decisions, such as controversial policy proposals made during U.S. state elections. Kriplean et al. [23] explored ways to promote active listening in web discussions by explicitly encouraging discussion members to summarize the points they heard. Kriplean et al. [21] studied the correctness of statements made in public deliberation and developed an on-demand fact-checking system. Zhang et al. [41] introduced recursive summarization of discussion trees to enable large scale discussions. Liu et al. [24] proposed a visualization technique to augment deliberation for multi-criteria decision making. Their results suggest that highlighting disagreement across multiple decision criteria can cause participants to align their

opinions for various reasons, from genuine consensus to appeasement. This finding reinforces the importance of designing deliberation procedures to minimize *groupthink*.

Most similar to our work, the MicroTalk workflow proposed by Drapeau et al. [7] focuses on argumentation within the microtask crowdsourcing setting. In MicroTalk, workers are prompted to provide justifications for their decisions, and an algorithm selects certain justifications (based on a metric of readability) to present them as counterarguments to other workers, triggering them to reconsider their decision. MicroTalk is based on an asynchronous model with no interactive back and forth discussion between workers. While we embed our work in a similar context as Drapeau et al., our focus is on understanding how synchronous, unfiltered group deliberation, task types and other characteristics impact resolvability, beyond just answer accuracy.

3 DELIBERATION WORKFLOW

We designed a workflow enabling crowdworkers to revisit and potentially resolve disagreements in text classification tasks through group discussions. The input to our workflow is a set of cases (e.g., text documents) to be classified. Each disagreement case either gets resolved through discussion or remains unresolved. The output of our workflow are multiple classification labels for each input case and its deliberation data consisting of structured information (i.e., original and reconsidered classification decisions, confidence levels, sources of disagreement, and evidence) and deliberation dialogues (e.g., arguments, explanations, and examples).

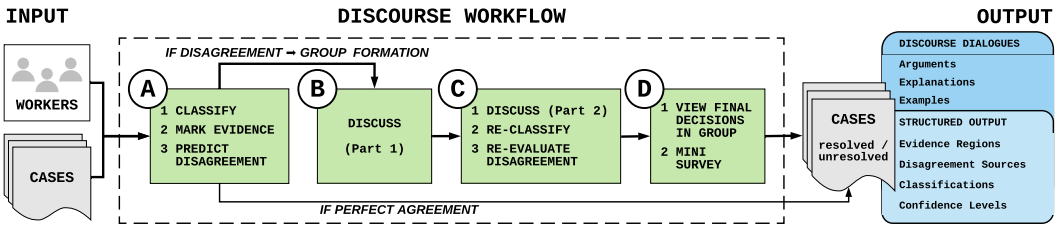

Fig. 1. Input, output and stages of the deliberation workflow implemented for this study. Each of the green blocks represents one of multiple microtasks in each of the four workflow stages A, B, C and D.

This workflow transforms input to output in four time-limited stages (Figures 1). Workflow stages A, B, C and D started at consecutive hours (e.g., 3pm, 4pm, 5pm, and 6pm), to ensure that all workers could complete one stage before collectively starting the next.

Stage A: Independent Classification. Workers independently performed 10 classification microtasks. In each task, workers were asked to read a text document, classify it into one of two categories, provide a confidence level for their decision, and highlight evidence to support their choice. Workers were then asked to predict the level of disagreement for the classification task by indicating whether they expect Substantial Agreement ("I expect most people to agree with me"), Half Agreement ("I expect only about half of the people to agree with me"), or Substantial Disagreement ("I expect most people to disagree with me"). If workers chose one of the latter two options, they were also asked to choose the sources of disagreement they anticipated. They could select multiple options from a list of six preset options as described in Table 1—Fuzzy Definition, Missing Context, Contradictory Evidence, Important Details, Expertise Needed, Subjective Case—covering a variety of common sources of disagreement from prior work and our pilot studies. They could provide their own rationale using an optional free-form field.

Between stage A and B, our system dynamically put crowdworkers into groups of three to deliberate on one or more cases, with the constraint that there was exactly one dissenter for each

Table 1. Preset choices for sources of disagreement.

Fuzzy Definition	Other people may have different definitions of [sarcasm / relation] in mind.
Missing Context	The text is ambiguous because of missing context (for example, [the identity of the product / some important information about the person or the place] is unknown).
Contrad. Evidence	The text contains some features that indicate [sarcasm / relation] is expressed and other features that indicate [absence of sarcasm / relation is not expressed].
Important Details	The text contains relevant details other people could easily miss.
Expertise Needed	Someone with more experience or expertise may see or understand something about the text that I don't.
Subjective Case	This is a case where a person's answer would depend heavily on their personal preferences and taste.
Other	Requiring free-form answer if selected

disagreement case per group. To ensure group heterogeneity, the group formation procedure was randomized, i.e., all workers had the same chance of being assigned to a group, and it was equally likely for them to be grouped together with either two random workers they disagreed with, or one with whom they agreed and one with whom they disagreed.

Stage B: Discussion Round 1. Workers joined their assigned groups to discuss their disagreement cases. For each case, workers were required to leave at least one comment in the group chat explaining why they had chosen their label for the given case. In addition, they could choose to highlight more parts of the text as evidence. The comments and associated evidence were recorded and shown to other workers in real time.

Stage C: Discussion Round 2 and Reconsideration. Workers collectively returned to review each other's comments and participate in a second round of discussion on the same disagreement cases. They were asked to further discuss the case by leaving at least one comment in the group chat, and optionally highlight more evidence. After providing at least one additional discussion comment, workers were individually prompted to reconsider their original classification decision and confidence level. To submit their final classification, workers could choose one of the two original class labels or a third option named "irresolvable". Finally, workers were also asked to re-evaluate what they considered the source of disagreement for the given case and group in light of the previous discussion. Workers were again given the six preset options (Table 1) and an optional free-form field, as well as the free-form answer they had provided earlier as the anticipated source of disagreement, if any. By providing the "irresolvable" option, we reduced the bias for consensus, and incentivized workers to change their answer only if they truly believed that their updated classification label was correct. Since we were interested in case resolvability, the reconsidered classification decisions were collected from all discussion members individually instead of enforcing the group to produce one joint decision. Importantly, workers did not see the updated answers of other group members before stage D to reduce opportunities for strategic voting.

Stage D: View Final Decisions. For each disagreement case, workers were presented with the final decision (i.e., either one of the class labels or "irresolvable") from each group member, and they were given a short open-ended survey on why they thought the disagreement was resolved or not resolved, as well as why they had changed or stuck to their original classification decision.

3.1 General Design Considerations

In the workflow design, we made several design decisions regarding the communication medium used for deliberation, ensuring stable pay/work ratios through filler tasks, and motivating workers to return for all four consecutive workflow stages despite the intermittent breaks.

Communication Medium. Our decision to use text (versus voice or video) as the communication medium for deliberation was motivated by prior research suggesting that written communication improves outcomes of consensus formation by avoiding bias from the tone of voice or a perceived lack of anonymity [3, 34].

Filler Tasks. If workers had less than 10 disagreement cases to revisit in stages B, C and D, they were asked to perform filler classification tasks (identical to the microtasks in stage A) to fill up the slots. This was to ensure that workers would not be incentivized to provide answers in stage A that were likely to agree with others just to reduce the number of disagreement cases they would have to process in subsequent stages. The data from the filler tasks were not used in our experiment.

Worker Retention. An important design consideration was the mechanism used to encourage workers to collectively return to stages B and C for real-time discussion. Through several pilot studies, we found a combination of monetary incentive and timed notifications to be a successful approach. As a monetary incentive, we paid fixed amounts for participation in each stage and an additional bonus for full participation in all stages. We used the MTurk API to send out reminder emails five minutes before the start of stages B, C and D, with a web link to join the next stage and a notice stating that the next stage could only be joined within three minutes of its start.

3.2 Interface

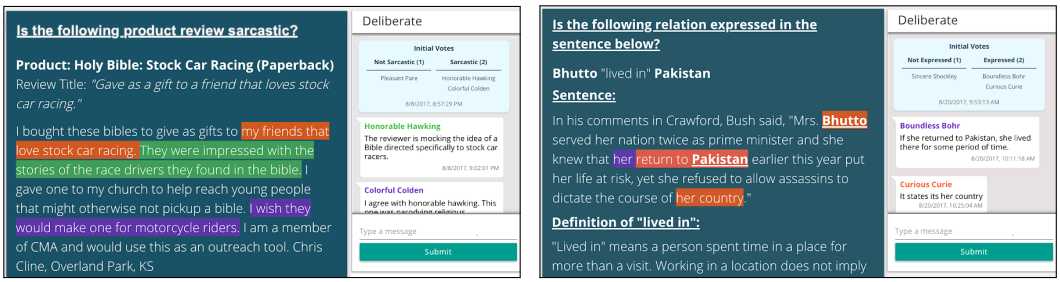

Fig. 2. Screenshots showing our worker deliberation interface paired with text classification tasks for sarcasm detection (left) and relation extraction (right). The text documents under discussion are shown on the left of each screenshot. Parts of the text are highlighted by different group members as evidence supporting their classification decisions. Written justifications are exchanged through a real-time chat component, with an embedded voting summary, shown on the right of each screenshot.

Our deliberation interface is general, i.e., it can be integrated into any task interface for text classification and requires only minor modifications for other data modalities like images. Figure 2 shows the interface in action. The interface displays the text document in question and attaches a chat box to the right, through which group members can exchange justifications in real time. Group members are randomly assigned friendly-sounding and easy-to-remember pseudonyms (e.g., 'Joyful Joliot' or 'Enthusiastic Easley') to allow users to address each other without having to reveal their identity. Finally, the interface enables users to highlight parts of the text document as evidence to support their argument. Highlights and pseudonyms are color-coded per deliberator.

4 EXPERIMENT

We conducted an experiment to investigate how deliberation affects the outcomes of crowdsourced text classification tasks. Here, we describe the task types, procedure, participant recruitment and payment associated with our study.

4.1 Task Types

We focus on two types of text classification tasks (Figure 2) with different degrees of inherent subjectivity. During our pilot studies, we also considered other task types (e.g., image, audio or video classification), but decided to defer those to future work due to the added complexity of synchronously highlighting evidence in such data modalities.

Sarcasm Task. For the first task type, we asked workers to label Amazon product reviews as sarcastic or not sarcastic, using the sarcasm detection data set published by Filatova [9]. We chose this task type and data set because Sarcasm detection is considered by Filatova et al. [9] as an inherently subjective task due to the "absence of a formal definition of sarcasm", an observation further supported empirically by low rates of inter-rater agreement in Filatova's data set. To generate a subset of cases for our experiment, we first identified all reviews for which Filatova [9] reported highest inter-rater disagreement (i.e., 3 sarcastic vs. 2 not sarcastic or vice versa) and then, from this subset, retained the 40 most compact cases based on word count.

Relation Task. For the second task type, workers were asked to indicate whether a certain semantic relation between a person and a place (e.g., "Nicolas Sarkozy *lived in* France", "Pavarotti *died in* Modena") was expressed in a given sentence. In contrast to the sarcasm detection task, this task is more well-defined and objective, as the ground truth data can be determined from the official label guidelines for the TAC KBP relations *LivedIn* and *DiedIn* as published by the Linguistic Data Consortium. We presented the corresponding relation definition to workers in each individual classification task (see right side of Figure 2 for an example) to explicitly make workers aware of the label guidelines. We used all 40 sentences from the data set used by Drapeau et al. [7], of which 25 have ground truth labels.

4.2 Procedure

Before the experiment, workers first filled out a pre-study questionnaire eliciting demographic information, including age group, gender, native language and self-rated proficiency in English. Participants then collectively stepped through the four consecutive stages of our workflow. They were free to close the browser tab or do other work after completing each stage and before the start of the next one, of which they were notified via email five minutes prior to start.

4.3 Participant Recruitment

We recruited 316 participants on Amazon Mechanical Turk, using workers from the US who had completed at least 500 tasks with a 90% acceptance rate. Based on the pre-study questionnaire, almost all of our workers are native English speakers (97%) and have high self-rated proficiency in English (96% and 4% selected the highest and second highest levels on a 5-point Likert scale). The distribution over age groups is: 18-25 (14%), 26-35 (44%), 36-45 (23%), 46-55 (13%) and 56+ (6%). About half of our workers (53%) are female.

4.4 Payment

Workers were paid US $1 for each stage that they completed. In addition, we paid a one-time completion bonus of US $2 to workers who completed all four stages. Each stage took workers around 15 minutes to complete, resulting in an approximate payment of US $6 per hour of work for participants who completed all four stages. Note that workers were free to close the browser tab or do other work after completing each stage and before the start of the next one. We incentivized workers to actively engage in discussion by offering an extra bonus of US $0.50 for each disagreement case in which all group members voted for the correct expert answer (which can be one of the two class labels, or "irresolvable") in stage C. As only few cases had an expert answer available, we paid the extra bonus to all groups in the end that reached consensus on one of the two target categories.

5 RESEARCH QUESTIONS AND HYPOTHESES

Our study aims to answer three research questions.

Q1: Why do annotators disagree with one another? We expect disagreement to arise due to different reasons in the two task types with different degrees of inherent subjectivity, where the target concepts are more versus less well-defined. Based on this intuition, we hypothesize that:

[H1a] Sources of disagreement differ significantly between the two task types.
[H1b] Annotators can predict levels of disagreement for individual cases better than random.

Q2: Under what circumstances can disagreement be resolved through worker deliberation? A variety of factors can contribute to the resolvability of a given case. First, the characteristics of a task and its associated sources of disagreement can play a role. For example, well-defined target categories may provide better grounding for convincing arguments, enabling groups to more easily come to a consensus. We hypothesize that:

[H2a] Sources of disagreement affect whether a case will be resolved.
[H2b] Task subjectivity affects whether a case will be resolved.

Second, the characteristics of the deliberation activities can influence whether a case is resolved. We hypothesize that:

[H2c] The extent to which members contributed equally affects case resolvability.

Third, we expect the amount of consensus in the label and overlap in highlighted evidence during the independent classification phase (i.e., stage A) to be predictive of a case's resolvability. We hypothesize that:

[H2d] The extent of the disagreement amongst group members affects resolvability.
[H2e] The amount of overlapping evidence between group members affects resolvability.

Q3: What impact does the deliberation workflow have on crowdsourcing outcomes and processes? The deliberation workflow, which encourages workers to consider diverse evidence and arguments, may have a positive effect on crowdsourcing outcomes and processes, such as improving the overall answer correctness and discouraging *groupthink* (i.e., the tendency of group members to blindly follow the majority). We hypothesize that:

[H3a] Worker deliberation improves the quality of the crowdsourced annotations.
[H3b] The probability that a case will be resolved in favour of the initial majority vote is similar to the probability that a case will not be resolved in favour of the initial majority vote.
[H3c] Sources of disagreement and the extent to which members contributed equally affects whether a case will be resolved correctly.

6 EXPERIMENTAL CONDITIONS

One of the goals in our study is to understand the effect of worker deliberation on the quality of crowdsourced annotations (H3a). To quantify the effects, we tested two additional variants of our workflow *without* the discussion component. Each participant was randomly assigned to one of the following conditions:

Disagree, Discuss and Reconsider (N=316): workers reconsider their position *after discussion* with other group members. This is our main condition testing our full deliberation workflow. If not otherwise noted, all results below are based on data from this condition.
Disagree and Reconsider (N=26): workers reconsider their position after they are shown group disagreement data, but *without* a discussion. This condition was added to isolate potential effects of showing workers information on who agreed vs. disagreed with them.
Reconsider Only (N=24): workers reconsider their position *without* being shown group disagreement data and *without* a discussion. This condition was added to identify any *learning* effects resulting from workers revisiting a case after labeling other cases.

The latter two conditions are used for hypothesis H3a only. In the results section for hypothesis H3a, we use the term **Baseline** for labels submitted during stage A (i.e., before any reconsideration), and otherwise refer to reconsidered labels submitted during workflow stage C.

7 DATA AND ANALYSIS

Data. For the analysis, the data include: (a) pre-study questionnaire data about demographics and language proficiency, (b) post-study questionnaire data, probing at workers' thoughts about and experiences with the deliberation process, (c) all the messages exchanged in the deliberation workflow stages, (d) the pre- and post-deliberation classification label and confidence for each worker, (e) the highlighted evidence for each case, (f) the sources of disagreement as anticipated by workers before discussion and re-evaluated by workers after discussion, (g) the anticipated resolvability of each case, and (h) the anticipated level of disagreement for each case.

Method. For each task type, we ran a *breadth* analysis on 40 text documents with up to 3 independent group discussions per document in order to identify the level of ambiguity for each case. This was followed by a *depth* analysis of the 10 most ambiguous cases with up to 16 independent group discussions per case, used to answer Q2 and Q3.

We analyzed the data using both quantitative (e.g., basic descriptive statistics, regression models) and qualitative analysis of worker responses. For the logistic regression models [25], we used the step-wise regression procedure to select the best possible combination of variables that could explain the dependent variable, based on improvements to the Akaike information criterion (AIC). We also test the goodness of fit of each model using the Hosmer-Lemeshow [16], Osius-Rojek [30] and Stukel [37] tests at significance level 0.05.

Filtering. Due to worker dropout between the four workflow stages, there were groups that became inactive during the deliberation process or had too few members remaining at the end. Out of all 316 participants, 78.2% completed all four stages, 3.8% and 4.8% dropped out after stages B and C respectively, and 13.2% completed only stage A. Possible explanations for the moderate dropout after stage A are that some workers might not have received or seen their email notifications in time (or not at all) to join stage B, or may simply not have been interested in returning to the same type of task in subsequent sessions. We excluded groups from the analysis if more than one member was inactive (i.e., they did not complete stages B and C), resulting in empty or single person groups, or where the *minority member* was inactive, leading to groups of two people sharing the same opinion. This resulted in two types of groups that were retained for the final analysis:

- **2 vs. 1** (unbalanced): all group members were active.
- **1 vs. 1** (balanced): one member dropped out, leaving two members with divergent opinions.

From a total number of 418 groups, we excluded 110 (26%) for the aforementioned reasons, resulting in 308 active groups retained for the analysis, of which 206 (67%) were unbalanced (2 vs. 1) and 102 (33%) were balanced (1 vs. 1) groups.

Data Set. The full data set is published at: https://github.com/crowd-deliberation/data.

8 RESULTS

Analysis based on descriptive statistics shows that there was more disagreement in the Sarcasm task than in the Relation task, confirming our premise that the Sarcasm task is more subjective. Specifically, we computed the level of label disagreement for each of the 40 cases per task type using the data from the *breadth* run. Label disagreement was represented as entropy:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

where p was the proportion of workers who chose the positive category (e.g., Sarcastic). Entropy values range from 0 to 1; higher values mean more disagreement (e.g., 50% choosing Sarcastic and 50% choosing Not Sarcastic) and lower values indicate less disagreement. Cases had mean entropy values of 0.85 ($SD = 0.22$) in the Sarcasm task and 0.61 ($SD = 0.32$) in the Relation task, indicating less disagreement overall in the Relation task. This difference is statistically significant under a two-sided t-test $t(70) = 3.79, p < 0.001$.

Table 2. Anticipated sources of disagreement (before discussion) where the proportions of workers are significantly different.

	Percentage of Participants	
Task	Relation	Sarcasm
Missing Context	**49%**	4%
Contrad. Evidence	18%	**43%**

Table 3. Re-evaluated sources of disagreement (after discussion) where the proportions of workers are significantly different.

	Percentage of Participants	
Task	Relation	Sarcasm
Fuzzy Definition	28%	**43%**
Missing Context	**24%**	4%
Contrad. Evidence	11%	**23%**

8.1 Q1: Why do annotators disagree with one another?

[H1a] To discover the sources of disagreement in each task, we analyzed the anticipated and re-evaluated sources of disagreement that workers provided before and after discussion. As workers were allowed to select multiple options, we use the simultaneous Pearson independence test [1], which takes into account correlation between options. Results (in Tables 2 and 3) show that certain sources of disagreement reported by the workers depend significantly on the task type, both, as anticipated before discussion ($\chi^2_S = 61.96, p < 0.001$) and as re-evaluated after the discussion ($\chi^2_S = 89.88, p < 0.001$). This result confirms our hypothesis H1a. A higher percentage of workers anticipated Missing Context to be a dominant source of disagreement in the Relation task (49%) than in the Sarcasm task (4%). Note that we included Missing Context to capture the *extent* to which it leads to disagreement, even though, following the standard TAC KBP guidelines for relation extraction, workers were instructed to avoid inferences based on missing information in the Relation task. Conversely, more workers anticipated Contradictory Evidence to be a dominant source of disagreement in the Sarcasm task (43%) than in the Relation task (18%). After discussion, while the same trends are observed, workers also identified Fuzzy Definition as an additional source of disagreement, which is more prominent in the Sarcasm task (43%) than the Relation task (28%).

[H1b] For Q1, we also investigated workers' ability to predict disagreement. Workers were asked to predict the level of disagreement before the discussion by choosing Substantial Disagreement, Half Agreement, or Substantial Agreement. We determined the ground truth "level of disagreement" for each case by running a two-sided proportion test to see if there is a 50/50 split in group opinions about the label. If this null hypothesis could not be rejected at significance level 0.05, we assumed the correct ground truth answer to be Half Agreement. Otherwise, the correct ground truth answer was based on the label chosen by majority vote. Based on the *depth* analysis data, 7 out of 10 cases in the Sarcasm task and 3 out of 10 cases in the Relation task resulted in Half Agreement.

Since very few workers predicted Substantial Disagreement (< 1% in each task type), this answer option was grouped together with Half Agreement (known as Disagreement from hereon) to allow for further inferential statistical tests. We measured workers' ability to predict the resulting two discrete levels Disagreement and Agreement (previously known as Substantial Agreement). For both task types, we computed workers' rate of being correct when predicting Disagreement or Agreement (*precision*), their rate of detecting all cases with Disagreement or Agreement (*recall*), as well as their overall prediction performance, in terms of *balanced accuracy*, a robust measure which accounts for class imbalance, defined as the average of the recall values for Disagreement and Agreement.

Figure 3 shows that workers' prediction performance was higher in the Relation task in terms of all the metrics, except for Disagreement precision. In other words, only the rate of being correct when predicting Disagreement was slightly higher in the Sarcasm task; for all other

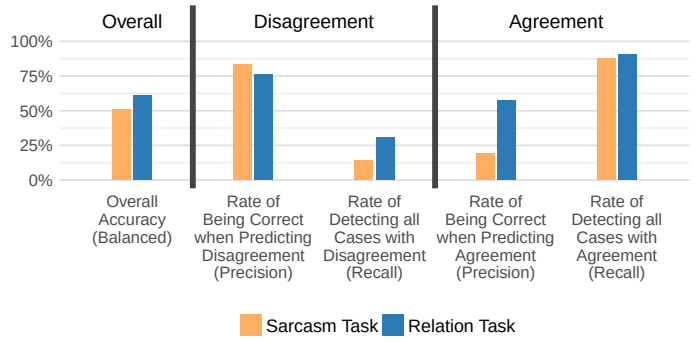

Fig. 3. Aggregate performance of our worker population at predicting disagreement/agreement by task type.

metrics, workers were more successful in the Relation task. In terms of the overall prediction performance, workers' accuracy was significantly better than random in the Relation task at 61% ($\chi^2(1, N = 716) = 158.63, p < 0.001$), whereas this was not the case for the Sarcasm task. This result partially confirms our hypothesis H1b. More interestingly, irrespective of the task type, when workers did predict Disagreement, their rates of being correct were higher than when predicting Agreement; but they were also generally less successful at detecting all cases with Disagreement than those with Agreement.

In summary, for Q1, we confirmed that sources of disagreement differ significantly between the two task types, identified Missing Context as a characteristic source of disagreement for the Relation task, and Contradictory Evidence and Fuzzy Definition for the Sarcasm task (H1a). In addition, we showed that workers can predict levels of disagreement significantly better than random in the Relation task (H1b).

8.2 Q2: Under what circumstances can disagreement be resolved through worker deliberation?

We analyzed the factors contributing to the resolution of disagreement using both quantitative and qualitative analyses of the questionnaire data. Group discussions were considered resolved if and only if all group members converged to one of the two target categories in their final classification.

For the quantitative analysis, a logistic regression model was used to discover factors that affect whether a case is resolved or not. The variables considered include, for each potential source of disagreement, the proportion of workers within a group who selected this source after discussion (H2a), the task type (H2b), various statistics capturing the amount of words contributed by group members (H2c), the level of initial consensus (H2d), and the amount of overlap between the highlighted pieces of evidence among the dissenting parties within a group (H2e), as measured by the Jaccard index, a statistic for the overlap between two sets. The pairwise interactions between the selected factors were excluded due to lack of statistical significance.[1]

[H2a] Results (Table 4) show that certain sources of disagreement—namely, Subjective Case, Fuzzy Definition, and Contradictory Evidence—decrease the probability of a case being resolved, partially confirming our hypothesis H2a.

In the post-study questionnaire, workers describe some cases as straightforward, requiring only a second glance to reach consensus ("After rereading, I realize the whole review is a joke and the

[1]Statistically significant results are reported as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

Table 4. Logistic model for understanding the likelihood of resolving a case.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*t*	*p-value*
Subjective Case	-2.16	0.83	-2.59	**
Fuzzy Definition	-1.45	0.53	-2.73	**
Contrad. Evidence	-1.48	0.65	-2.27	*
# Words Min/Max	-1.49	0.67	-2.24	*
Group 2 vs. 1	-0.67	0.33	-2.04	*
Sarcasm Task	0.61	0.36	1.69	

positive points are sarcastic."), while other cases are ambiguous or confusing due to contradictory evidence (e.g., the text contains features of both sarcasm and non-sarcasm) and missing context.

[H2b] We speculated that task subjectivity affects whether a case will be resolved. Results are mixed. Our step-wise regression procedure showed that that disagreements in the Sarcasm task are more likely, although statistically insignificant, to be resolved than in the Relation task (H2b). In contrast to that, in the post-study questionnaire, workers in the Relation task mentioned that *well-defined category definitions* helped them resolve disagreement ("We were able to refer to the definition of LivedIn. That made the answer clear."), while the "lack of instruction on what constitutes a sarcastic review" was considered a barrier to resolving cases in the Sarcasm task.

[H2c] We hypothesize that the extent to which members contributed equally predicts whether the case will be resolved. Our model selected # Words Min/Max, i.e., the proportion of words contributed by the least active and the most active contributors within a group, as a significant predictor variable, confirming hypothesis H2c. An increase in this proportion decreases the likelihood of a case being resolved. In other words, if members contributed equally, cases were significantly less likely to be resolved than if some members contributed substantially more than others.

Qualitative analysis also shows that interactions between members of the discussion groups can influence the final outcomes. Workers said that disagreement was resolved by clarifying the *task* ("Members were becoming clearer on interpretation of the task."), providing *examples* ("Using examples, we were able to persuade the group member to change their opinion."), using *evidence* ("We were able to point out things in the sentence that were overlooked by others."), or pointing out *false assumptions* ("Others pointed out things that some of us had assumed in the sentence and changed our opinions.") Many workers identified the *quality of deliberation activity* itself as the main driver for reaching consensus in an argumentative manner, e.g., "People that didn't agree listened to the arguments and were willing to change their mind to sarcastic.") On the other hand, workers also reported some group-related factors that hindered the resolution of disagreement, including divergent, but equally valid *interpretations* ("I think it depends on how people view sarcasm.") and the *lack of communication* ("[My group was] not continuing a conversation. Responding with one word responses does not solve anything.")

[H2d] A consensus level of 2 vs. 1 (i.e., two workers agree, one worker disagrees) also decreases the likelihood of a case being resolved compared to the consensus level of 1 vs. 1, confirming our hypothesis H2d.

[H2e] Finally, the overlap in evidence among group members was not selected as a relevant factor for the optimal model fit, leading us to reject hypothesis H2e.

To summarize the findings for Q2, we found various factors that affect whether a case is resolved or not, including some sources of disagreement (H2a), task type (H2b), the degree to which deliberation

activity is balanced within a group (H2c) and the level of initial consensus (H2d). Other factors like overlap in evidence highlighted by dissenting parties (H2e) had no significant effect on resolvability. Finally, our analysis shows that worker characteristics (such as age and personality) can have some influence on the way they deliberate (e.g., their overall tendency to revise their position), which in turn can play a role in resolving a case.

8.3 Q3: What impact does the deliberation workflow have on crowdsourcing outcomes and processes?

[H3a] To evaluate whether worker deliberation improves crowdsourcing outcomes, we compared workers' answer correctness between our three experimental conditions **Reconsider Only**, **Disagree and Reconsider**, and **Disagree, Discuss and Reconsider** and our **Baseline** (i.e., original labels submitted in stage A). This part of the analysis is restricted to the 25 cases from the *breadth* run, where we have ground truth to assess correctness.

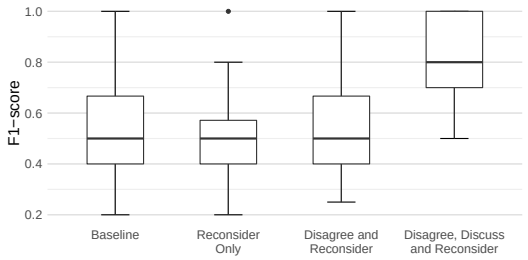

Fig. 4. Individual workers' answer quality in the Relation task across different reconsideration workflows.

Correctness was measured by F1-score (i.e., the harmonic mean of precision and recall) of *individual* workers. We did not aggregrate labels across multiple workers as our workflow did not contain a requirement for shared unanimous group decisions, but instead incentivized independent reconsideration from individual workers in all conditions. Assessments that were revised to "irresolvable" were excluded because they were neither right nor wrong. Overall, correctness was significantly higher in **Disagree, Discuss and Reconsider**, with no significant differences between the other three conditions. Figure 4 provides a visual comparison. The statistical significance of these differences was confirmed by a one-way analysis of variance ($F(3, 115) = 6.42, p < 0.001$), followed by pairwise comparisons with Holm-Bonferroni correction. The pairwise comparisons confirmed that the **Disagree, Discuss and Reconsider** workflow resulted in significantly higher F1-scores than **Baseline** ($t(64) = -3.44, p < 0.01$), **Reconsider Only** ($t(44) = -4.38, p < 0.001$), and **Disagree and Reconsider** ($t(25) = -3.02, p < 0.05$). There were no detectable differences among the other pairings. These results confirm our hypothesis H3a that worker deliberation improves the quality of the crowdsourced annotations. Furthermore, this result suggests that the improvement in correctness is not due to learning effects or knowledge about group disagreement data, but due to the actual discussion process.

[H3b] For Q3, we were also interested in whether our proposed deliberation workflow discourages *groupthink*, an undesirable effect where discussion members tend to agree with the original majority answer within a group. For this question, we performed a close-up analysis on all groups with a composition of 2 vs. 1, and analyzed whether the proportion of cases resolved in favour of the original majority vote was similar to the proportion of cases not resolved in favour of the original majority vote. We used a two-sided proportion test which confirmed our null hypothesis H3b that both outcomes were equally likely, $\chi^2(1, N = 136) = 0.60, p = 0.44$. In other words,

unbalanced discussion groups were equally likely to converge to the original majority opinion as they were to achieve the opposite outcome, i.e., leave the case unresolved or converge to the original minority opinion. Our quantitative results provide evidence that our proposed deliberation workflow effectively discourages *groupthink*.

[H3c] For Q2, we identified factors that contribute to the resolvability of a case. For hypothesis H3c, we investigate whether some of the factors considered in Q2, e.g., sources of disagreement and the extent to which members contributed equally, also predict whether or not a case will be resolved *correctly*. We performed an analysis on the subset of cases for which ground truth was available and which were resolved in the deliberation process. We used a similar step-wise logistic regression procedure as for Q2, defining the *correctness* of the final consensus label as the dependent variable and including the same set of predictor variables. Results (Table 5) show that certain sources of disagreement made the *correct* resolution of a case more likely (Missing Context) or less likely (Expertise Needed). The extent to which members contributed equally (# Words Min/Max) was a negative predictor for correct resolution. These results partially confirm our hypothesis H3c.

Table 5. Logistic model for understanding the likelihood of resolving a case correctly.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*t*	*p-value*
Expertise Needed	-3.89	1.95	-1.99	*
Missing Context	2.91	1.42	2.05	*
# Words Min/Max	-3.43	1.73	-1.99	*

To summarize, we provide evidence showing that worker deliberation significantly improves the quality of crowdsourced annotations (H3a), while discouraging undesirable behaviour such as *groupthink* (H3b), and that certain sources of disagreement and the extent to which members contribute equally help predict the correctness of the final resolution (H3c). To recapitulate the primary results of this study, we provide a high-level summary of each hypothesis and the degree to which it was supported in Table 6.

9 DISCUSSION

In this work, we studied how deliberation affects resolvability and answer accuracy, and how deliberation outcomes depend on task subjectivity. Our results demonstrate that legitimate reasons for disagreement can vary by task, and worker deliberation can help resolve some of these cases. Importantly, we identified several factors (such as the level of initial consensus, the amount and quality of deliberation activities, and sources of disagreement) that contribute to case resolvability. We reinforced the finding that deliberation can increase answer accuracy and the importance of verbal discussion in this process. Finally, our deliberation workflow discouraged undesirable behaviour, such as *groupthink*. Our anonymized data set is publicly available for research reuse.

9.1 Generalizability and Applications

Our empirical findings are based on experiments using two specific task types, one subjective sarcasm detection task, and one objective (person-to-place) relation extraction task. Caution is warranted for translating these results to a broader set of classification tasks and data modalities. However, the deliberation workflow we propoosed is general, and can be applied to classification tasks involving other data types (e.g., images, videos, time series data), given minor modifications to the interface to facilitate annotation of evidence in other modalities. In some complex scenarios

Table 6. High-level summary of all hypotheses and the degrees to which they were supported.

	Q1: Why do annotators disagree with one another?	
H1a	Sources of disagreement differ by task type.	Supported (***)
H1b	Annotators can predict disagreement levels.	Supported for Relation task (***)
	Q2: Under what circumstances can disagreement be resolved through worker deliberation?	
H2a	Sources of disagreement affect resolvability.	Supported for Subj. Case (**), Fuzzy Def. (**) and Contrad. Evid. (*)
H2b	Task subjectivity affects resolvability.	Partially supported
H2c	Extent of equal contribution affects resolvability.	Supported (*)
H2d	Level of initial consensus affects resolvability.	Supported (*)
H2e	Amount of overlap in evidence affects resolvability.	Not supported
	Q3: What impact does the deliberation workflow have on crowdsourcing outcomes and processes?	
H3a	Worker deliberation improves answer correctness.	Supported (***)
H3b	Groupthink is discouraged by our deliberation incentives.	Supported
H3c	Sources of disagreement and the extent of equal contribution affect whether cases get resolved correctly.	Supported for Expertise Needed (*), Missing Context (*) and the extent of equal contribution (*)

requiring hierarchical decision processes, our workflow will need to be modified. For example, in image classification, individual image features might be disambiguated first before resolving disagreement on the image level. Our results revolve around atomic classification problems in paid crowdwork, and it is expected that deliberation processes in more complex settings would give rise to more complex dynamics (e.g., circular disagreement in sequential classification).

In general, deliberation workflows like ours can be particularly useful in domains where inter-rater disagreement is rather the norm than exception, e.g., medicine [4, 8, 33]. Beyond enabling workers to discuss and reconsider ambiguous cases, our deliberation workflow collects rich information from human annotators (e.g., confidence levels, arguments, assumptions, examples, inferences, relevant features from the data, and meta information about sources of disagreement) that can be used to teach both humans and machines. For example, prior work has shown that asking annotators to highlight relevant features in input data (i.e., words in a text document or regions in an image) can improve machine inference in sentiment analysis tasks [35, 38–40] and visual category learning [5]. More generally, we posit that future efforts towards *interpretable* machine learning can use data produced through group discussion to analyze, replicate and mimic human decision making processes. While some of the less structured deliberation output (e.g., verbal arguments) may not yet be fully parsed by automated methods, human learners could significantly benefit from edge-case examples coming with discussions and highlighted features from more experienced annotators. Existing efforts to optimize or harness workers' ability to learn complex tasks [6, 31] could thus leverage data produced through worker deliberation.

Another outcome of deliberation could be the refinement and disambiguation of annotation guidelines or category definitions for expert tasks. For example, scoring manuals in the field of medical imaging undergo regular revisions to increase inter-rater reliability [4, 33], a procedure which could benefit from structured and web-based deliberation workflows.

9.2 Design Considerations for Classification Tasks

Redundant Labeling. Our results demonstrate that workers can predict levels of inter-rater disagreement significantly better than random for certain task types. Prior work has shown that being able to predict answer diversity can reduce cost because fewer annotators are needed when answer agreement is expected [14]. Enlisting human capabilities to predict answer agreement incurs minimal extra cost because it requires one additional human response for each data object.

Task Interfaces. Our results show that certain sources of disagreement are more likely to occur for some task types than for others. In classification tasks (like our Relation task) that require annotators to make decisions solely based on the information provided in the data, the interface could remind annotators who indicate "Missing Context" as the reason for anticipating disagreement to minimize ungrounded assumptions about any latent contextual information.

Deliberation Workflow. We found a variety of factors that affect case resolvability. Deliberation workflows for crowdsourced classification tasks should therefore be equipped with incentive mechanisms to reduce or strengthen the effect of certain factors in a task-specific manner. For example, in task types where objectivity is possible and the goal is to find one correct answer, deliberation procedures should have incentives to reduce the impact of undesirable behaviour (e.g., stubbornness or lack of care) and undesirable group dynamics (like *groupthink* or lack of communication) on the final discussion outcome. Deliberation workflows for more subjective classification tasks where the goal is to uncover multiple divergent, but equally valid interpretations of the task and data (e.g., sentiment analysis, relevance rating, text translation) should incentivize group members to be assertive about their interpretation, and change their assessment only under certain conditions (e.g., when false assumptions or illogical conclusions are pointed out).

Various deliberation systems have been proposed in complex domains like public deliberation [10, 22, 23], on-demand fact checking [21], and knowledge base generation [41]. While our study is primarily embedded in the domain of paid crowd work, our empirical findings and some aspects of our workflow design may be informative for deliberation systems in general. For example, our insight that sources of disagreement, when captured in structured form, can help predict case resolvability (H2a) could be leveraged to categorize debates and streamline consensus building. Our finding that equal contribution among discussion members seemed to negatively affect case resolvability (H2c) *and* final answer correctness (H3c) was surprising as balanced contribution is often considered beneficial for fruitful discussion. This finding could inspire future investigations into the balance between active and passive forms of contribution to deliberation, related to work by Kriplean et al. [23] on active listening in web discussions. Another potentially counter-intuitive finding of our work is the fact that unbalanced (2 vs. 1) groups were *less* likely to resolve a case than balanced (1 vs. 1) groups (H2d). To our knowledge, our study is the first to investigate the effect of initial consensus level on case resolvability by having multiple independent groups of size two or three discuss the same case in crowdsourced classification tasks. While one may expect that unbalanced groups converge faster, one possible explanation for our observation is the added complexity of communication and coordination among three versus two group members. We showed that a perceived lack of expert knowledge was associated with resolving disagreements incorrectly (H3c). This result has interesting connections to prior work on integrating on-demand fact checking into public deliberation [21], suggesting that crowdsourcing workflows could benefit from similar approaches for on-demand provision of expertise.

9.3 Limitations and Future Work

We studied the effects of deliberation in the context of two binary classification tasks and small groups consisting of only two or three members. In practice, many classification problems have more than two classes and discussion groups can also be larger. Future work can investigate new methods for scaling to more complex classification problems and more complex group structures.

A practical limitation of our workflow is its reliance on multiple rounds of synchronous communication and potential attrition between stages. Dropout was highest after the first stage in our workflow, suggesting promising future research on incentives for workers to return for multiple consecutive sessions, or on systems that initiate discussions immediately as disagreement arises before workers leave the platform.

Our deliberation workflow enables workers to consider alternative views on ambiguous cases through discussion, and produces useful data such as arguments, examples, and evidence. However, deliberation also incurs a cost in terms of time. A promising area for future work is to develop techniques to improve the *efficiency* of the deliberation process [7] and to characterize the cost-benefit of using real-time deliberation in task workflows. Another direction for future work is to explore other deliberation incentives beyond monetary compensation, including peer-based reputation systems for constructive deliberation [11]. In more complicated incentive schemes, moderators could be rewarded for establishing a balanced and constructive deliberation among group members.

10 CONCLUSION

This work contributes novel insights into the circumstances and outcomes of worker deliberation for handling inter-rater disagreement in crowdsourced text classification tasks with varying degrees of inherent subjectivity. Based on a custom-designed workflow for real-time worker deliberation, we investigated the impact of various factors on the probability that a disagreement among small groups of crowdworkers will be resolved through synchronous group discussion. Our results suggest that the reasons for and level of the initial disagreement, the amount and quality of deliberation activities, as well as the task and case characteristics play a role for resolvability. To encourage future work in the field of worker deliberation, we publish our data set including all original classifications, discussion comments, text highlights, and revised positions from crowdworkers. Future work includes developing and validating new deliberation protocols and demonstrating how the information produced by worker deliberation can be used to train both humans and machines.

ACKNOWLEDGEMENT

This work was funded by NSERC CHRP (CHRP 478468-15) and CIHR CHRP (CPG-140200).

REFERENCES

[1] Christopher R. Bilder and Thomas M. Loughin. 2004. Testing for Marginal Independence between Two Categorical Variables with Multiple Responses. *Biometrics* 60, 1 (3 2004), 241–248. DOI: http://dx.doi.org/10.1111/j.0006-341X.2004.00147.x

[2] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM, ACM Press, New York, New York, USA, 2334–2346. DOI: http://dx.doi.org/10.1145/3025453.3026044

[3] Norman Dalkey and Olaf Helmer. 1963. An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science* 9, 3 (4 1963), 458–467. DOI: http://dx.doi.org/10.1287/mnsc.9.3.458

[4] Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, Bernd Saletu, Andrea Schmidt, and Georg Dorffner. 2009. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research* 18, 1 (3 2009), 74–84. DOI: http://dx.doi.org/10.1111/j.1365-2869.2008.00700.x

[5] Jeff Donahue and Kristen Grauman. 2011. Annotator rationales for visual recognition. In *2011 International Conference on Computer Vision*. IEEE, 1395–1402. DOI: http://dx.doi.org/10.1109/ICCV.2011.6126394

[6] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 2623–2634. DOI: http://dx.doi.org/10.1145/2858036.2858268

[7] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[8] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (7 2018), 1–20. DOI: http://dx.doi.org/10.1145/3152889

[9] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation - LREC '12*, Nicoletta Calzolari, Khalid

Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), 392–398. DOI: http://dx.doi.org/10.1.1.681.7735

[10] Deen G. Freelon, Travis Kriplean, Jonathan Morgan, W. Lance Bennett, and Alan Borning. 2012. Facilitating Diverse Political Engagement with the Living Voters Guide. *Journal of Information Technology & Politics* 9, 3 (7 2012), 279–297. DOI: http://dx.doi.org/10.1080/19331681.2012.665755

[11] Snehalkumar (Neil) S. Gaikwad, Mark Whiting, Karolina Ziulkoski, Alipta Ballav, Aaron Gilbee, Senadhipathige S. Niranga, Vibhor Sehgal, Jasmine Lin, Leonardy Kristianto, Angela Richmond-Fuller, Jeff Regino, Durim Morina, Nalin Chhibber, Dinesh Majeti, Sachin Sharma, Kamila Mananova, Dinesh Dhakal, William Dai, Victoria Purynova, Samarth Sandeep, Varshine Chandrakanthan, Tejas Sarma, Adam Ginzberg, Sekandar Matin, Ahmed Nasser, Rohit Nistala, Alexander Stolzoff, Kristy Milland, Vinayak Mathur, Rajan Vaish, Michael S. Bernstein, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, and Sharon Zhou. 2016. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*. ACM Press, New York, New York, USA, 625–637. DOI: http://dx.doi.org/10.1145/2984511.2984542

[12] Luciana Garbayo. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539 (2014), 35–45. http://link.springer.com/10.1007/978-3-319-04280-0_5

[13] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*. https://arxiv.org/pdf/1703.08774.pdf

[14] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM, ACM Press, New York, New York, USA, 3511–3522. DOI: http://dx.doi.org/10.1145/3025453.3025781

[15] Francis T. Hartman and Andrew Baldwin. 1995. Using Technology to Improve Delphi Method. *Journal of Computing in Civil Engineering* 9, 4 (10 1995), 244–249. DOI: http://dx.doi.org/10.1061/(ASCE)0887-3801(1995)9:4(244)

[16] David W. Hosmer and Stanley Lemesbow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* 9, 10 (1980), 1043–1069. DOI: http://dx.doi.org/10.1080/03610928008827941

[17] Alan M. Jones. 1973. Victims of Groupthink: A Psychological Study of Foreign Policy Decisions and Fiascoes. *The ANNALS of the American Academy of Political and Social Science* 407, 1 (5 1973), 179–180. DOI: http://dx.doi.org/10.1177/000271627340700115

[18] Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*. ACM Press, New York, New York, USA, 1635–1646. DOI: http://dx.doi.org/10.1145/2818048.2820016

[19] Sara Kiesler and Lee Sproull. 1992. Group decision making and communication technology. *Organizational Behavior and Human Decision Processes* 52, 1 (6 1992), 96–123. DOI: http://dx.doi.org/10.1016/0749-5978(92)90047-B

[20] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (3 2018). DOI: http://dx.doi.org/10.1016/j.ophtha.2018.01.034

[21] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, New York, New York, USA, 1188–1199. DOI: http://dx.doi.org/10.1145/2531602.2531677

[22] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012a. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, New York, New York, USA, 265. DOI: http://dx.doi.org/10.1145/2145204.2145249

[23] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012b. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, New York, New York, USA, 1559. DOI: http://dx.doi.org/10.1145/2207676.2208621

[24] Weichen Liu, Sijia Xiao, Jacob T Browne, Ming Yang, and Steven P Dow. 2018. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *ACM Transactions on Social Computing* 1, 1 (1 2018), 4:1–4:26. DOI: http://dx.doi.org/10.1145/3159649

[25] Peter McCullagh and John Nelder. 1989. *Generalized Linear Models* (2 ed.). Chapman & Hall/CRC.

[26] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[27] Jeryl L. Mumpower and Thomas R. Stewart. 1996. Expert Judgement and Expert Disagreement. *Thinking & Reasoning*

2, 2-3 (7 1996), 191–212. DOI: http://dx.doi.org/10.1080/135467896394500

[28] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* (1 2018). DOI: http://dx.doi.org/10.1038/s41562-017-0273-4

[29] Charlan Nemeth. 1977. Interactions Between Jurors as a Function of Majority vs. Unanimity Decision Rules. *Journal of Applied Social Psychology* 7, 1 (3 1977), 38–56. DOI: http://dx.doi.org/10.1111/j.1559-1816.1977.tb02416.x

[30] Gerhard Osius and Dieter Rojek. 1992. Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *J. Amer. Statist. Assoc.* 87, 420 (12 1992), 1145–1152. DOI: http://dx.doi.org/10.1080/01621459.1992.10476271

[31] Shengying Pan, Kate Larson, Joshua Bradshaw, and Edith Law. 2016. Dynamic Task Allocation Algorithm for Hiring Workers that Learn. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*. New York, 3825–3831.

[32] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (7 2017). http://arxiv.org/abs/1707.01836

[33] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (1 2013). DOI: http://dx.doi.org/10.5664/jcsm.2350

[34] Harold Sackman. 1974. *Delphi assessment: Expert opinion, forecasting, and group process.* Technical Report. RAND CORP SANTA MONICA CA.

[35] Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active Learning with Rationales for Text Classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*.

[36] Miriam Solomon. 2006. Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy* 44, S1 (3 2006), 28–42. DOI: http://dx.doi.org/10.1111/j.2041-6962.2006.tb00028.x

[37] Thérèse A. Stukel. 1988. Generalized Logistic Models. *J. Amer. Statist. Assoc.* 83, 402 (6 1988), 426–431. DOI: http://dx.doi.org/10.1080/01621459.1988.10478613

[38] Ainur Yessenalina, Yejin Choi, and Claire Cardie. 2010. Automatically Generating Annotator Rationales to Improve Sentiment Classification. In *Proceedings of the ACL 2010 Conference Short Papers (ACLShort '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 336–341. http://dl.acm.org/citation.cfm?id=1858842.1858904

[39] Omar F. Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. 260–267.

[40] Omar F. Zaidan, Jason Eisner, and Christine D. Piatko. 2008. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NIPS 2008 Workshop on Cost Sensitive Learning*.

[41] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 2082–2096. DOI: http://dx.doi.org/10.1145/2998181.2998235

Received April 2018; revised July 2018; accepted September 2018