# Debiased Label Aggregation for Subjective Crowdsourcing Tasks

SHAUN WALLACE*, Brown University, USA
TIANYUAN CAI*, Adobe Inc., USA
BRENDAN LE, Brown University, USA
LUIS A. LEIVA*, University of Luxembourg, Luxembourg

Human Intelligence Tasks (HITs) allow people to collect and curate labeled data from multiple annotators. Then labels are often aggregated to create an annotated dataset suitable for supervised machine learning tasks. The most popular label aggregation method is majority voting, where each item in the dataset is assigned the most common label from the annotators. This approach is optimal when annotators are unbiased domain experts. In this paper, we propose Debiased Label Aggregation (DLA) an alternative method for label aggregation in *subjective* HITs, where cross-annotator agreement varies. DLA leverages user voting behavior patterns to weight labels. Our experiments show that DLA outperforms majority voting in several performance metrics; e.g. a percentage increase of 20 points in the $F_1$ measure before data augmentation, and a percentage increase of 35 points in the same measure after data augmentation. Since DLA is deceptively simple, we hope it will help researchers to tackle subjective labeling tasks.

CCS Concepts: • **Information systems** → *Crowdsourcing*; • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; • **Computing methodologies** → *Supervised learning*.

Additional Key Words and Phrases: Label aggregation; Voting bias; Supervised learning; Crowdsourcing

**ACM Reference Format:**
Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A. Leiva. 2022. Debiased Label Aggregation for Subjective Crowdsourcing Tasks. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3491101.3519614

## 1 INTRODUCTION

Machine Learning (ML) has become ubiquitous in Human-Computer Interaction (HCI), from gesture recognition [51] and text entry [27] to chatbots [81] and user interface design [82]. Supervised learning is one of the most popular approaches to train ML models, where the task is to infer a function from labeled training data and then use it to make predictions on new, unseen data. Researchers and practitioners often use Human Intelligence Tasks (HITs) in crowdsourcing settings to collect and curate labeled data. Generally, in this setting, multiple crowdworkers label each item in a dataset, and then labels are aggregated so that each item has exactly one label. While some researchers argue that supervised ML models should train on *soft labels* that represent the distribution of crowdworkers' annotations [42, 77], most classification tasks must make a single discrete decision [25, 83], for example recognizing a gesture shortcut from regular handwriting [39] or understanding if a review is positive or negative [15].

*Denotes equal contribution.


Manuscript submitted to ACM

Ensuring that item labels are accurate is a challenging and ongoing research problem [73] because they are considered ground-truth data in supervised ML classifiers. Simply put, if an ML model sees the wrong label for a given item during training, it will not learn to classify similar items when deployed in a real-world application properly. These real-world applications often use predictions to provide valuable recommendations to users. Therefore suboptimal recommendations from misclassifications can result in users perceiving the application as unusable [85].

Crowdsourcing platforms such as Amazon Mechanical Turk[1] or Prolific[2] provide an efficient and relatively inexpensive approach to obtaining substantial amounts of labeled data using HITs. Ideally, multiple crowdworkers should assess every item in a dataset to produce more accurate labels [12, 72, 84]. Then, majority voting is commonly used for label aggregation by researchers and practitioners because it is computationally straightforward to implement [16, 30, 34, 70]. Majority voting assigns each item the most common label made by the crowdworkers who labeled that item.

If we assume that every annotator is a true expert in the labeling task's subject, prior research shows that majority voting is the optimal decision rule for label aggregation [66]. However, this assumption does not always hold in practice since crowdworkers are diverse among their prior background, cognitive abilities, and experience, thus affecting their labeling behavior even for the same piece of data [9, 17]. These traits of labeling behavior are particularly critical when labeling *subjective* data since complete label agreement towards a given item happens very rarely [25]. For example, while it can be trivial to identify whether a picture contains some object, subjective tasks can become incredibly difficult, such as determining if a sketch is inspirational [74] or identifying if a human face is real or artificial [3].

In this paper, we propose Debiased Label Aggregation (DLA), an alternative method to majority voting in subjective HITs. We focus on the binary labeling problem, a popular scenario in crowdsourcing settings [12, 56, 61, 78, 79]. However, DLA can be extended to multi-class labeling tasks without loss of generalizability, as explained later. Also, binary labeling tasks are prevalent among researchers when studying subjective tasks such as assessing sentence toxicity [2], disambiguation of satellite imagery [63], and determining gender bias in course selection [48], among others.

DLA leverages patterns in users' voting behavior across multiple items within a dataset to improve the quality of the aggregated labels. DLA integrates a user's voting history across a dataset to reweight their binary labels to reduce the overall user bias when aggregating labels to create ground-truth labels. We study and compare DLA against majority voting on an public real-world dataset of highly subjective data [74]. Taken together, our experiments suggest that DLA shows better characterization of the labeled items and results in significant improvements over majority voting based on several performance metrics; see Figure 2. As shown in section 3, DLA is deceptively simple to implement and so we hope that researchers and practitioners will integrate DLA into their label aggregation process when dealing with subjective HITs. Our software implementation is publicly available.[3]

## 2 RELATED WORK

We focus our review of previous work on human biases and labeling of ambiguous and subjective data, as they are the core research areas in this paper. We survey previous research covering how annotator bias affect ML models and understanding annotator biases for subjective data.

[1] https://www.mturk.com/
[2] https://prolific.co/
[3] https://github.com/open-hci-research/debiased-label-aggregation

### 2.1 Can Annotator Bias Affect Model Outcomes?

Hiring a single crowdworker to label data is not only unscalable but also insufficient due to an individual's biases and cognitive abilities [33, 67]. Previous work tried to overcome this by hiring multiple crowdworkers to label a single item and used majority voting (i.e., the most common label) to determine the ground-truth labels [30, 75].

Prior work shows that a majority voting label aggregation scheme requires a very large number of labels to overcome voting bias, even for simple tasks [24, 33, 61]. One underlying assumption is that an annotator's level of expertise might affect their bias [64, 76]. However, prior work shows this assumption does not hold [26, 33]. Several prior research efforts show different label aggregations methods are necessary when working with ambiguous data [23, 61]. Prior work also shows these alternative aggregation methods can help to reduce the total number of labels needed to train an ML model compared to majority voting [40, 44]. Our DLA method uses an individual's voting history to help break ties, reduce the total number of votes needed, and does not require explicit assessment of an individual's domain expertise.

While prior work shows promise in labeling subjective data, inferring truth from labels is still an open problem [86]. There are many issues to overcome, ranging from task design, data quality, and human biases [5, 33, 67, 86]. A recent effort by Gordon et al. [25] introduced a label transformation method to align ML classification metrics with user-facing performance measures. Their method is based on the observation that current label aggregation methods (in particular the standard majority voting method) ignore disagreement between people. Cabrera et al. developed a method to debias labels that are incorrect because of poor quality data [5]. While poor data quality is an issue, other research shows that annotators are not necessarily wrong; they have differing opinions when labeling ambiguous data [1]. DLA focuses on subjective data where an annotator is never considered wrong. This simple viewpoint can reduce the technical complexity to determine ground-truth labels, whereas other research focuses on extracting extra demographic information, evaluating expertise, or requiring additional tasks from an annotator [64, 86]. Thus, it is possible to use DLA on any existing dataset with as little information as the unique identifiers for items and their annotators.

### 2.2 Understanding Annotator Biases for Subjective Data

The decision to label subjective data to generate ground-truth datasets is difficult. Previous research shows that different individuals can generate different labels for the same piece of data [9, 17]. Dumitrache et al. used Knowlton's triangle of reference to help explain why this occurs [49] and identified three sources of disagreement in crowdsourced labeling tasks, namely: *Sign* (the data itself is ambiguous or subjective); *Conception* (the annotator has different perspectives); and *Referent* (the labeling task is ambiguous or poorly designed). Kairam and Heer studied the triangle of reference for labeling tasks and found each to affect ground-truth labels [35]. Our work uses a public dataset; therefore, we believe "referent" is not pertinent to this paper's research goals. Thus we rely on the "sign" to explain item difficulty and "conception" to explain differing opinions among annotators. We believe a highly subjective dataset will elicit both causes.

Chung et al. were motivated by previous work focusing on ambiguous labeling efforts to investigate how to collect multiple labels efficiently for subjective data [9]. Their efforts focused on improving disagreement due to "referent." An area of future work they identified is to improve the estimation capability of the annotators by showing them other examples to label and the collective answer distributions for this data. While this idea would increase the time and costs to generate labels, our work goes a step beyond this idea by accounting for an individual annotator's biases and the data's subjectivity. Disagreement among annotators can be caused by different perspectives or expertise [38, 60]. Our work builds on these prior ideas to potentially improve pre-existing and future labels to create high-quality datasets.

Finally, we found that a large body of previous work trying to mitigate voting bias behavior has largely ignored downstream tasks, such as using the labeled data for classification and recommendation (i.e., they have focused solely on alternative label aggregation methods and have evaluated them with experts as oracles) without training any ML model [1, 2, 6, 9, 16, 17, 25, 33, 35].

For any method to be adopted in practice, we believe it is necessary to create and use the aggregated labels as ground-truth data in ML models. In this paper, we show that DLA shows better characterization of the labeled items and results in significant improvements over majority voting on the basis of several performance metrics.

## 3 METHOD

As previously mentioned, the most common method for label aggregation in HITs is using a simple majority vote [30, 36]. This method can ensure data consistency where multiple crowdworkers (usually an odd number, to avoid ties) label the same item. Then the most common answer is considered a close reflection of the ground-truth label for that item. However, this approach omits additional information gathered from crowdworkers to improve the quality of the aggregated labels. Critically, a majority voting approach does not account for user bias and expertise. In the following, we propose DLA, our novel label aggregation method that has proved to work remarkably well in subjective HITs.

### 3.1 Dataset and Research Background

To develop DLA, we use a public dataset of sketches with binary labels (inspirational/non-inspirational) collected during six voluntary non-graded sketching activities in a university-level UI/UX classroom [74]. We chose this dataset because it contains highly subjective data and has the raw data necessary to train a ML classifier.[4] Labeling inspiration is subjective because it relies on both "sign" to explain item difficulty and "conception" to explain differing opinions among annotators [49]. Therefore, a robust label aggregation approach is necessary to elicit more accurate labels to train an ML classifier to predict if a sketch is inspirational or non-inspirational.


Fig. 1. Samples from the Sketchy dataset [74]. Sketches range from simplistic to complex, and from less to more complete.

Binary labels were collected while students were sketching. They were allowed to "peek" (i.e., view) their peers' sketches and see them develop in real-time. While peeking, participants answered either "yes" or "no" to the question: "Will you change your sketch based on what you see in this sketch?" Figure 1 depicts some examples of the submitted sketches. The authors who ran the sketching tasks [74] used this question as a proxy for in-the-moment inspiration

[4] Many other public datasets we examined only provide crowdsourcing labels but not the raw data.

based on prior research [11, 50, 69]. This binary inspiration question is more subjective than questions in traditional labeling tasks.

We observed that participants were likely to label a sketch as non-inspirational 89.4% of the time. Based on this information, we will refer to the non-inspirational and inspirational sketches respectively as the majority ("+") and minority ("-") classes in our classification task; see next section. Also, we observed that some participants tended to label sketches as inspirational more often than others, which indicates a highly skewed voting behavior across participants that should be factored in.

### 3.2 Weighting Annotator Labels to Debias Voting Behavior

When labeling subjective data, it is wise to consider the voting behavior of each user and use that to weight their vote. Let $x$ be an item that received $n$ votes $\mathcal{S} = (s_1, s_2, ..., s_n), \quad s_j \in \{-, +\}$, where $j$ correspond to a vote, and where $s_j = -$ corresponds to a vote belonging to the minority class, and $s_j = +$ corresponds to a vote belonging to the majority class. Note that a multi-class classification task can be decomposed as multiple binary "one-vs-all" classification tasks [22, 57], in which case the target class (the "one") can be seen as the minority class and the rest (the "all") as the majority class, respectively. The final class label can be assigned according to simple majority voting:

$$C(x) = \underset{i \in \{-,+\}}{\arg\max} \sum_{j=1}^{n} \mathbb{1}_{\{s_j = i\}} \tag{1}$$

where $\mathbb{1}_{\{s_j=i\}}$ denotes the indicator function, which equals to 1 if the $j$th vote is equal to $i$ and 0 otherwise. For example, given votes $\mathcal{S} = (-, -, +)$ that three independent crowdworkers have assigned to some item $x$, using majority voting we would classify $x$ as "class -", since $C(x) = \arg\max[\sum \mathbb{1}_{\{s_j=-\}}, \sum \mathbb{1}_{\{s_j=+\}}] = \arg\max[2, 1] = -$. As noted, the voting scheme in (1) is naive since it assumes all users have the same voting behavior, which we denote as their prior voting probability. Therefore, we should compensate how $C(x)$ is computed to account for users that score one class more often over the other class.

Our method calculates the weights $w_j$ for each item vote $s_j \in \mathcal{S}$ separately, based on the voting history per user. Let $p_j = \frac{\sum +}{n_j}$ be the user bias, i.e., the proportion of "class +" votes from the user who cast vote $s_j$, where $n_j$ is the total number of votes cast for that user. The weight assigned to that user is given by:

$$w_j = \begin{cases} p_j, & \text{if } s_j = - \\ 1 - p_j, & \text{if } s_j = + \end{cases} \tag{2}$$

As noted, the weight for a positive vote (i.e., a vote belonging to the majority class) becomes the proportion of votes from the user who cast that negative vote. And vice versa for negative votes. For users who mostly vote positive, their positive votes will be weighted lower and their negative votes higher, and vice versa for users who vote primarily negative. Using this approach, the average weighted vote from any particular user becomes exactly $1/c$, where $c$ is the number of classes, thus removing any potential voting bias for label aggregation. Using Equation 2, we propose the following expression to debias voting behavior:

$$\widetilde{C}(x) = \underset{i \in \{-,+\}}{\arg\max} \sum_{j=1}^{m} w_j \mathbb{1}_{\{s_j = i\}} \tag{3}$$

Continuing with the previous example of $\mathcal{S} = (-, -, +)$, suppose now that our three crowdworkers have the following voting proportions, which determine the following weights:

$$p_1 = 0.2 \rightarrow w_1 = 0.2; \quad p_2 = 0.2 \rightarrow w_2 = 0.2; \quad p_3 = 0.3 \rightarrow w_3 = 0.7$$

We would now classify $x$ as "class +" since

$$\begin{aligned} \widetilde{C}(x) &= \arg\max[\sum w_j \mathbb{1}_{\{s_j=-\}}, \sum w_j \mathbb{1}_{\{s_j=+\}}] \\ &= \arg\max[0.2 + 0.2, 0.7] \\ &= +. \end{aligned}$$

To improve DLA further, we introduce a simple normalization procedure to account for some users having more "voting history" than others, whose vote should count more [79]. Therefore, we can consider the number of votes of each user $n_j$ as a multiplier factor in Equation 3:

$$\widetilde{C}(x) = \underset{i \in \{-,+\}}{\arg\max} \sum_{j=1}^{m} n_j w_j \mathbb{1}_{\{s_j=i\}} \tag{4}$$

Equation 4 is recommended if there is a large variability in the number of votes per user, otherwise, $n_j$ would be similar for all users and therefore Equation 4 would become Equation 3 in practice.

## 4 EVALUATION

We validate our approach by training the same ML model on the labels aggregated by majority voting, Equation 3, and Equation 4, respectively. A model trained on labels of higher quality should yield better classification performance.

### 4.1 Model Training and Data Preprocessing

We built an Extreme Gradient Boosting (XGBoost) classifier, which is an ensemble of weak decision trees [8] and performed cross-validation to evaluate the effectiveness of the ML model.

We note that neural networks are a common choice for ML practitioners. We did try several convolutional neural network configurations but they did not achieve good performance in this dataset. Thus we chose a high-performance feature-based statistical classifier[5].

We configured our XGBoost classifier with 500 estimators and a maximum tree depth of 10. These parameters were estimated after a stratified 10-fold grid search sampling. It is vital to estimate the model configuration carefully for many reasons. For example, shallow trees usually have poor performance because they capture too few details of the problem. Also, deeper trees generally capture too many details of the problem and may overfit the training dataset, thus limiting the ability to make useful predictions on new data.

We selected general-purpose features to represent a given sketch, based on Wallace et al. [74], which they based on stroke recognition research literature and related areas [10, 21, 31, 32, 41, 43, 45, 52, 53, 58, 62, 68, 80]. Table 1 shows the feature set. As explained below, we then perform feature selection to remove highly correlated features.

We perform feature whitening and oversampling as a data preprocessing step. On the one hand, feature whitening (scale normalization) transforms the values for each computed feature in the $[0, 1]$ range to prevent features with greater numeric ranges from dominating those with smaller numeric ranges. On the other hand, a dataset with unbalanced classes may perform poorly because it will see one class more often during training. So it will tend to predict the majority

[5]The XGBoost classifier is among the top winners in ML competitions like Kaggle[6] and has achieved state-of-the-art results in a variety of tasks [8].

| **Global Features** | | **Local Features** | |
|---|---|---|---|
| Aspect ratio (1,2) | Num. strokes (1,3) | Avg. point angles (1,2,3) | SD point distances (1,2,3) |
| Bounding box angle (1,2,3) | Num. fitted strokes | Avg. point diff (1) | SD point diff (1,2) |
| Bounding box area | Num. points (1) | Avg. point distances (1,2,3) | SD point angles (1,2) |
| Bounding box length (1,2,3) | Num. fitted points (1) | Sum point angles (2,3) | Cosine initial angle (1,3) |
| Convex hull area (2) | Num. corners | Sum point differences (1,2,3) | Sine initial angle (2) |
| Path length (1) | | Sum point distances (2,3) | Cosine end angle (1,2,3) |
| Strokes entropy (1,2,3) | | Sum squared point angles (1,2,3) | Sine end angle (1,2,3) |

Table 1. List of features for sketch classification. The final feature set for each model (denoted in brackets) is determined after recursive feature elimination. Models trained with labels from: majority voting (1), Equation 3 (2), and Equation 4 (3).

class more often. Therefore, we perform SMOTE oversampling on the training data, which interpolates new samples from the minority class [7]. Both whitening and oversampling are standard preprocessing techniques in supervised ML.

Not all features are relevant for recognition. Generally, a smaller feature set is preferred since statistically weak features prevent generalization to unseen data, introduce noise in the model, and increase training time. Therefore, feature selection becomes indispensable. We used recursive feature elimination (RFE), a procedure that selects features by considering increasingly smaller sets of features in a cross-validation loop [13]. First, an estimator is trained on the initial set of features, and then, based on the statistical importance of each feature ($p$-values), the least important features are removed (if $p < .05$). This procedure repeats until there are no more irrelevant features.

Previous work has shown that any label aggregation method really benefits from a large number of votes per item [66]. Therefore, aimed at investigating the scalability of DLA, we tested the effect of increasing the number of collected votes. For this, we repeated the previous experimental setup with data augmentation, by duplicating each user vote 5 times. In the following section we report the results of both set of experiments.

### 4.2 Performance Measures

Classification accuracy is not the only relevant performance metric to assess an ML classifier. Area Under the ROC curve (AUC), for example, helps to determine the discriminatory power of any classifier [55]. Further, we report the classic Precision and Recall metrics as well. Precision is the number of true positives (i.e., the number of inspirational sketches correctly predicted) divided by the total number of sketches predicted as inspirational, and Recall is defined as the number of true positives divided by the total number of actual inspirational sketches. We also report the $F_1$ measure, which is the (evenly weighted) harmonic mean of Precision and Recall, and the Jaccard index [20], which estimates the likelihood of a sketch being inspirational if it is not correctly classified as non-inspirational. In sum, we report a variety of classification metrics to gain a full understanding of model performance.

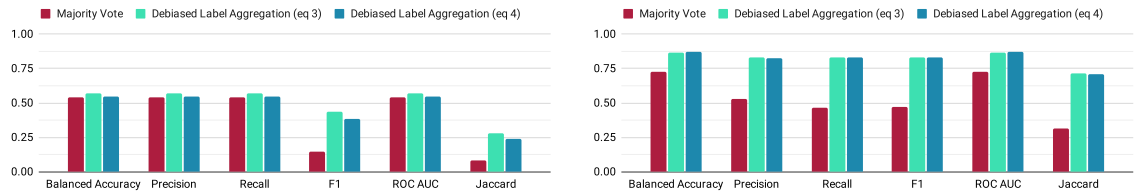


Fig. 2. Classification performance of the same ML classifier trained on three different label aggregation approaches, both before (left) and after data agumentation (right). Note: 95% confidence intervals are all below 2.5%, so they are omitted.

### 4.3 Results

Comparing the classification performance of models trained on majority voting and our two proposed DLA variants, we find that considering the user's bias when weighting the votes produces labels that improve overall classification performance; see Figure 2. Notice that the improvements attributed to DLA outperform the baseline performance of majority voting by a large margin, especially when more user votes are available. This improvement after data augmentation was expected, since each item has more information about the agreement distribution of user votes. Therefore, the three label aggregation approaches we studied led to better classification performance.

Table 2 shows the proportion of label changes per method, both before and after data augmentation. We can note that DLA is more robust than majority voting, since the differences in most of the classification metrics are more apparent when more user votes are considered for label aggregation.

| **Method** | **No. original items** | | **No. augmented items** | |
|---|---|---|---|---|
| | **class - (%)** | **class + (%)** | **class - (%)** | **class + (%)** |
| Majority Voting | 386 (98.4) | 6 (1.6) | 2005 (96.3) | 77 (3.7) |
| DLA Equation 3 | 265 (64.4) | 146 (35.6) | 1353 (65.7) | 706 (34.3) |
| DLA Equation 4 | 274 (66.6) | 137 (33.4) | 1291 (62.7) | 768 (37.3) |

Table 2. Upon label debiasing, we noticed an increase in the number of positive labels as compared with majority voting, both before and after data augmentation.

Our results suggest that labeling subjective items in a crowdsourcing setting should depart from the classic majority voting approach and adopt more informed methods like the one we propose in this paper. In summary, aggregating labels while considering user voting behavior is a promising approach to creating higher quality labels in subjective crowdsourcing tasks.

## 5 DISCUSSION

Creating ML models for predicting subjective behavior is challenging, as users have personal biases when performing labeling tasks. This idea of asking someone for a subjective rating or vote applies to many areas outside of ML and HCI. For example, someone would find it hard to trust a friend's favorable restaurant recommendation if that friend believes over 90% of restaurants are good. However, if that same friend said a restaurant was bad, that recommendation would carry more weight or impact. Our work has noticed this same behavior applied to users labeling sketches as inspirational or not, a highly subjective labeling task.

We developed a straightforward but powerful approach to label aggregation based on users' voting behavior for subjective tasks. Our DLA method leads to significant improvements in an ML classifier that can predict whether a sketch is inspirational. This contribution shows possibilities for how future researchers can develop predictive models to answer difficult, subjective questions.

### 5.1 Recommendations for Applying DLA

When collecting a new dataset to generate ground-truth labels for various tasks, such as ML model training or statistical analysis, it can be challenging to know which user-provided labels to trust and what to ignore. Research into label aggregation with disagreements is currently under-researched [73]. In the past, to generate high-quality data collection

efforts, crowdsourcing researchers relied on the annotators to resolve disagreements among themselves or attempted to improve task instructions [4, 6, 17, 25, 46, 59]. Some prior work has sought to employ these methods in real-time during an active data collection effort. We hesitate to suggest that DLA applies during an active data collection effort because our method relies on voting history to improve outcomes. The more votes a user makes, the better we can understand their voting patterns. We recommend the usage of our method once data collection is complete. Furthermore, DLA can complement prior methods, such as improving task design [4], discussed to ensure that the dataset and labels are of the highest quality possible.

One line of research to ensure better quality labels and data focuses on detecting vandalism or intentionally bad behavior by annotators. Detecting vandalism is a common researched area with numerous proven methods to identify and discard intentionally incorrect data [14, 29, 54, 65, 75]. Compared to these prior methods, in the subjective labeling tasks we studied, an individual annotator with a different opinion than the majority does not necessarily mean they are wrong [1, 9, 17–19, 35]. They might have different viewpoints and experiences than the majority, which affects their labeling decisions. To reiterate, DLA should be applied later in the data evaluation process. This will ensure paid crowdworkers will receive fair compensation regardless of their viewpoints [71].

### 5.2 Limitations and Future Work

Our work does not seek to replace psychometric and survey design research for understanding individuals' shifting biases over time. We assume that labels are generated in a finite amount of time by a small number of crowdworkers. This assumption is reasonable and aligns with prior work, as it is common for a participant in a crowdsourcing setting to label some data on a budget and then for researchers to evaluate these contributions post-task. Future extensions to DLA could evaluate user biases across different tasks longitudinally to understand how biases evolve, for example.

Our paper has analyzed a single dataset, as we found it challenging to find public subjective datasets containing raw data (in addition to the collected crowdsourcing labels) to create an ML classifier. Previous work has utilized synthetic data to compensate for the lack of suitable datasets [5, 14, 34, 61]. We hope the community will continue studying subjective labeling and prediction tasks; for example, detecting emotions from facial expressions [37] or political views from tweets [28, 47]. Another future research direction is to investigate if DLA will apply to non-subjective data. A first step should be to estimate annotators' voting skewness and decide whether majority voting should be applied. If the voting distribution is heavily biased, then applying DLA makes sense.

## 6 CONCLUSION

In crowdsourcing scenarios, majority voting is not an appropriate label aggregation approach for subjective labeling tasks since it ignores (and is affected by) the users' skewed voting behavior. We have proposed DLA, a new label aggregation method to elicit information from ambiguous data sources, where user subjectivity is likely to impact the final labels' quality. This paper has focused on labeling inspirational sketches, serving as a prequel to expand DLA to other datasets. As previously shown, DLA is deceptively simple, and so we hope that researchers and practitioners will adopt it.

## ACKNOWLEDGMENTS

This work was supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001. This work is also supported by the National Science Foundation under Grant No. IIS-1552663.

## REFERENCES


[1] Lora Aroyo, Anca Dumitrache, Oana Inel, Zoltán Szlávik, Benjamin Timmermans, and Chris Welty. 2019. Crowdsourcing Inclusivity: Dealing with Diversity of Opinions, Perspectives and Ambiguity in Annotated Data. In *Proc. WWW Companion*. 1294–1295.

[2] Agathe Balayn and Alessandro Bozzon. 2019. Designing Evaluations of Machine Learning Models for Subjective Inference: The Case of Sentence Toxicity. In *REAIS Workshop a HCOMP*.

[3] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a deep generative model. In *Proc. ECCV*. Springer, 351–369.

[4] Jonathan Bragg and Daniel S Weld. 2018. Sprout: Crowd-powered task design for crowdsourcing. In *Proc. UIST*. 165–176.

[5] Guillermo F Cabrera, Christopher J Miller, and Jeff Schneider. 2014. Systematic labeling bias: De-biasing where everyone is wrong. In *Proc. ICPR*. 4417–4422.

[6] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proc. CHI*. 2334–2346.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* 16 (2002), 321–357.

[8] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. KDD*. 785–794.

[9] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 1–25.

[10] Gennaro Costagliola, Vincenzo Deufemia, Giuseppe Polese, and Michele Risi. 2004. A parsing technique for sketch recognition systems. In *Proc. VL/HCC*. 19–26.

[11] Arthur J Cropley. 1997. Fostering creativity in the classroom: General principles. *The creativity research handbook* 1, 84.114 (1997), 1–46.

[12] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. 2013. Aggregating crowdsourced binary ratings. In *Proc. WWW*. 285–294.

[13] Burcu F. Darst, Kristen C. Malecki, and Corinne D. Engelman. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics* 19 (2018).

[14] Luca De Alfaro, Vassilis Polychronopoulos, and Michael Shavlovsky. 2015. Reliable aggregation of boolean crowdsourced tasks. In *Proc. HCOMP*.

[15] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proc. CHI*. 1–14.

[16] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proc. KDD*. 259–268.

[17] Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *Proc. ESWC*. 701–710.

[18] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proc. HCOMP*.

[19] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing ground truth for medical relation extraction. *ACM Trans. Interact. Intell. Syst.* 8, 2 (2018).

[20] Sam Fletcher and Md Zahidul Islam. 2018. Comparing sets of patterns with the Jaccard index. *Australas. J. Inf. Syst.* 22 (2018).

[21] Clive Frankish, Richard Hull, and Pam Morgan. 1995. Recognition accuracy and user acceptance of pen interfaces. In *Proc. CHI*, Vol. 95. 503–510.

[22] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit.* 44, 8 (2011), 1761–1776.

[23] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Trans. Image Process.* 26, 6 (2017), 2825–2838.

[24] W. Gao, L. Wang, Y.-F. Li, and Z.-H. Zhou. 2016. Risk minimization in the presence of label noise. In *Proc. AAAI*. 1575–1581.

[25] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proc. CHI*. 1–14.

[26] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proc. AAAI*.

[27] Andrew Hard, Chloé M Kiddon, Daniel Ramage, Francoise Beaufays, Hubert Eichner, Kanishka Rao, Rajiv Mathews, and Sean Augenstein. 2018. Federated Learning for Mobile Keyboard Prediction. In *arXiv:1811.03604*.

[28] Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. 2018. Machine learning-based sentiment analysis for twitter accounts. *Appl. Math. Comput.* 23, 1 (2018), 11.

[29] Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. 2015. Towards vandalism detection in knowledge bases: Corpus construction and analysis. In *Proc. SIGIE*. 831–834.

[30] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. 2013. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Math. Comput. Modell.* 57, 11-12 (2013), 2918–2932.

[31] Jason I. Hong and James A. Landay. 2000. SATIN: A Toolkit for Informal Ink-Based Applications. In *Proc. UIST*. 63–72.

[32] Bing Quan Huang, YB Zhang, and Mohand Tahar Kechadi. 2007. Preprocessing techniques for online handwriting recognition. In *Proc. ISDA*. 793–800.

[33] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proc. CHI*. 1–12.

[34] Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Min. Knowl. Discov.* 28, 2 (2014), 402–441.

[35] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proc. CSCW*. 1637–1648.

[36] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Proc. NeurIPS*. 1953–1961.

[37] Eugene Laksana, Tadas Baltrušaitis, Louis-Philippe Morency, and John P Pestian. 2017. Investigating facial behavior indicators of suicidal ideation. In *Proc. FG*. 770–777.

[38] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D O'Keefe, and Walter S Lasecki. 2018. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proc. ACM Hum. Comput. Interact.* 2, CSCW (2018), 1–23.

[39] Luis A. Leiva, Vicent Alabau, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2014. Context-Aware Gestures for Mixed-Initiative Text Editing UIs. *Interact. Comput.* 27, 6 (2014).

[40] Christopher H Lin, Mausam Mausam, and Daniel S Weld. 2012. Crowdsourcing control: Moving beyond multiple choice. In *Proc. AAAI Workshops*.

[41] James S. Lipscomb. 1991. A trainable gesture recognizer. *Pattern Recognit.* 24, 9 (1991), 895–907.

[42] Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019. Learning to predict population-level label distributions. In *Proc. WWW Companion*. 1111–1120.

[43] Wenyin Liu. 2003. On-line graphics recognition: State-of-the-art. In *International Workshop on Graphics Recognition*. 291–304.

[44] Robert Tyler Loftin, James MacGlashan, Bei Peng, Matthew E Taylor, Michael L Littman, Jeff Huang, and David L Roberts. 2014. A Strategy-Aware Technique for Learning Behaviors from Discrete Human Feedback.. In *Proc. AAAI*. 937–943.

[45] A Chris Long Jr, James A Landay, Lawrence A Rowe, and Joseph Michiels. 2000. Visual similarity of pen gestures. In *Proc. CHI*. 360–367.

[46] VK Chaithanya Manam and Alexander J Quinn. 2018. Wingit: Efficient refinement of unclear task instructions. In *Proc. HCOMP*.

[47] Diana Maynard and Adam Funk. 2011. Automatic detection of political opinions in tweets. In *Proc. ESWC*. 88–99.

[48] Danaë Metaxa-Kakavouli, Kelly Wang, James A Landay, and Jeff Hancock. 2018. Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proc. CHI*. 1–6.

[49] Charles Kay Ogden and Ivor Armstrong Richards. 1923. The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. *Nature* 111, 566 (1923).

[50] Victoria C Oleynick, Todd M Thrash, Michael C LeFew, Emil G Moldovan, and Paul D Kieffaber. 2014. The scientific study of inspiration in the creative process: challenges and opportunities. *Front. Hum. Neurosci.* 8 (2014), 436.

[51] Sameera Palipana, Dariush Salami, Luis A. Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021).

[52] Brandon Paulson and Tracy Hammond. 2008. Paleosketch: accurate primitive sketch recognition and beautification. In *Proc. IUI*. 1–10.

[53] Brandon Paulson, Pankaj Rajan, Pedro Davalos, Ricardo Gutierrez-Osuna, and Tracy Hammond. 2008. What!?! no Rubine features?: using geometric-based features to produce normalized confidence values for sketch recognition. In *HCC Workshop: Sketch Tools for Diagramming*. 57–63.

[54] Martin Potthast. 2010. Crowdsourcing a Wikipedia vandalism corpus. In *Proc. SIGIR*. 789–790.

[55] D.M.W. Powers. 2011. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *J. Mach. Learn. Technol.* 2, 1 (2011).

[56] V. C. Raykar, Y. Shipeng, L. H. Zhao, G . H. Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning from crowds. *J. Mach. Learn. Res.* 11 (2010), 1297–1322.

[57] Ryan Rifkin and Aldebaro Klautau. 2004. In Defense of One-Vs-All Classification. *J. Mach. Learn. Res.* 5 (2004), 101–141.

[58] Dean Rubine. 1991. Specifying Gestures by Example. *Proc. SIGGRAPH* 25, 4 (1991), 329–337.

[59] 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 (2018), 1–19.

[60] Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Wang, and Brent Hecht. 2015. Turkers, Scholars," Arafat" and" Peace" Cultural Communities and Algorithmic Gold Standards. In *Proc. CSCW*. 826–838.

[61] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. KDD*. 614–622.

[62] Beat Signer, Ueli Kurmann, and M Norrie. 2007. iGesture: a general gesture recognition framework. In *Proc. ICDAR*, Vol. 2. 954–958.

[63] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring Ground Truth from Subjective Labelling of Venus Images. In *Proc. NeurIPS*. 1085–1092.

[64] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP*. 254–263.

[65] Yu Suzuki and Satoshi Nakamura. 2016. Assessing the quality of Wikipedia editors through crowdsourcing. In *Proc. WWW*. 1001–1006.

[66] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. 2019. Domain-Weighted Majority Voting for Crowdsourcing. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 1 (2019), 163–174.

[67] Fangna Tao, Liangxiao Jiang, and Chaoqun Li. 2020. Label similarity-based weighted soft majority voting and pairing for crowdsourcing. *Knowl. Inf. Syst.* 62 (2020), 2521–2538.

[68] Charles C. Tappert, Ching Y. Suen, and Toru Wakahara. 1990. The state of the art in online handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 12 (1990), 787–808.

[69] Todd M Thrash, Laura A Maruskin, Scott E Cassidy, James W Fryer, and Richard M Ryan. 2010. Mediating between the muse and the masses: Inspiration and the actualization of creative ideas. *J. Pers. Soc. Psych.* 98, 3 (2010), 469.

[70] T. Tian, J. Zhu, and Y. Qiaoben. 2019. Max-margin majority voting for learning from crowds. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 10 (2019), 2480–2494.

[71] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 1–26.

[72] Jinzheng Tu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, and Maozu Guo. 2018. Multi-label answer aggregation based on joint matrix factorization. In *Proc. ICDM*. IEEE, 517–526.

[73] Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 Task 12: Learning with Disagreements. In *ACL Workshop on Semantic Evaluation*.

[74] Shaun Wallace, Brendan Le, Luis A Leiva, Aman Haq, Ari Kintisch, Gabrielle Bufrem, Linda Chang, and Jeff Huang. 2020. Sketchy: Drawing Inspiration from the Crowd. *Proc. ACM Hum. Comput. Interact.* 4, CSCW2 (2020), 1–27.

[75] Shaun Wallace, Alexandra Papoutsaki, Neilly H Tan, Hua Guo, and Jeff Huang. 2021. Case Studies on the Motivation and Performance of Contributors Who Verify and Maintain In-Flux Tabular Datasets. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 1–25.

[76] Shaun Wallace, Lucy Van Kleunen, Marianne Aubin-Le Quere, Abraham Peterkin, Yirui Huang, and Jeff Huang. 2017. Drafty: Enlisting Users to be Editors who Maintain Structured Data. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. 187–196.

[77] Jing Wang and Xin Geng. 2019. Classification with Label Distribution Learning. In *Proc. IJCAI*. 3712–3718.

[78] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. In *Proc. NeurIPS*. 2424–2432.

[79] 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 *Proc. NeurIPS*. 2035–2043.

[80] Jin Xiangyu, Liu Wenyin, Sun Jianyong, and Zhengxing Sun. 2002. On-line graphics recognition. In *Proc. Pacific Graphics*. 256–264.

[81] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proc. CHI*. 3506–3510.

[82] Arianna Yuan and Yang Li. 2020. Modeling Human Visual Search Performance on Realistic Webpages Using Analytical and Deep Learning Methods. In *Proc. CHI*. 1–12.

[83] Biqiao Zhang, Georg Essl, and Emily Mower Provost. 2017. Predicting the distribution of emotion perception: capturing inter-rater variability. In *Proc. ICMI*. 51–59.

[84] Hao Zhang, Liangxiao Jiang, and Wenqiang Xu. 2019. Multiple Noisy Label Distribution Propagation for Crowdsourcing. In *Proc. IJCAI*. 1473–1479.

[85] Qian Zhao, F Maxwell Harper, Gediminas Adomavicius, and Joseph A Konstan. 2018. Explicit or implicit feedback? Engagement or satisfaction? A field experiment on machine-learning-based recommender systems. In *Proc. SAC*. 1331–1340.

[86] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endowment* 10, 5 (2017), 541–552.