Journal of Artificial Intelligence Research 61 (2018) 363-405 Submitted 10/2017; published 02/2018

Actively Estimating Crowd Annotation Consensus

Yunus Emre Kara YUNUS.KARA@BOUN.EDU.TR
Department of Computer Engineering, Bogazici University
TR-34342, Bebek, Istanbul, Turkey

Gaye Genc GAYE.GENC@BOUN.EDU.TR
Department of Computer Engineering, Bogazici University
TR-34342, Bebek, Istanbul, Turkey

Oya Aran ARANOYA@GMAIL.COM
Independent Researcher

Lale Akarun AKARUN@BOUN.EDU.TR
Department of Computer Engineering, Bogazici University
TR-34342, Bebek, Istanbul, Turkey

Abstract

The rapid growth of storage capacity and processing power has caused machine learning applications to increasingly rely on using immense amounts of labeled data. It has become more important than ever to have fast and inexpensive ways to annotate vast amounts of data. With the emergence of crowdsourcing services, the research direction has gravitated toward putting the wisdom of crowds to better use. Unfortunately, spammers and inattentive annotators pose a threat to the quality and trustworthiness of the consensus. Thus, high quality consensus estimation from crowd annotated data requires a meticulous choice of the candidate annotator and the sample in need of a new annotation. Due to time and budget limitations, it is of utmost importance that this choice is carried out while the annotation collection is in progress. We call this process *active crowd-labeling*. To this end, we propose an active crowd-labeling approach for actively estimating consensus from continuous-valued crowd annotations. Our method is based on annotator models with unknown parameters, and Bayesian inference is employed to reach a consensus in the form of ordinal, binary, or continuous values. We introduce ranking functions for choosing the candidate annotator and sample pair for requesting an annotation. In addition, we propose a penalizing method for preventing annotator domination, investigate the explore-exploit trade-off for incorporating new annotators into the system, and study the effects of inducing a stopping criterion based on consensus quality. We also introduce the crowd-labeled Head Pose Annotations datasets. Experimental results on the benchmark datasets used in the literature and the Head Pose Annotations datasets suggest that our method provides high-quality consensus by using as few as one fifth of the annotations ($\sim 80\%$ cost reduction), thereby providing a budget and time-sensitive solution to the crowd-labeling problem.

1. Introduction

In the machine learning domain, labeled datasets are valuable commodities. Computing resources have increased exponentially for two decades, driving machine learning toward big data applications. The introduction of the ImageNet database (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009), a large crowd-labeled dataset, and the success of deep neural network methods have further pushed the research direction toward the use of large datasets. This popularity has resulted in the introduction of many large crowd-labeled datasets such as the recently introduced Open Images dataset (Krasin et al., 2017).

Providing ground truth labels for large datasets often proves to be excessively time consuming. Thus, researchers tend to outsource the labeling process, especially for the aforementioned large datasets. However, employing expert labelers is expensive. Crowdsourcing the labeling process is a cost-effective and fast method to solve this problem, especially when expertise is not necessarily required.

Crowd-labeling is the process of collecting annotations from crowds and using them for estimating consensus values to be used as labels. However, inattentive annotators and spammers reduce the quality of consensuses. Although there are numerous methods in the literature that deal with the low quality annotations, most are effective only after the annotation process is completed. At this point, valuable time and money are already spent. Therefore, it is of utmost importance to observe and understand the behavior of the annotators early on in the annotation process and improve the quality of consensuses.

The classical use of crowd-labeling is analogous to a careless shopper who buys excessively without proper planning and ends up throwing away their purchase when the product is of low quality or unneeded. In contrast, imagine that the researcher is a meticulous shopper with limited time and money. The most important questions on their mind would be: What am I in need of purchasing and which vendor should I purchase it from? Applying this reasoning to the crowd-labeling problem calls for a smarter solution and active learning is the remedy to this problem. The general idea of active learning can be applied to the crowd-labeling problem in terms of choosing which annotation to incorporate into the annotation pool. In this work, the process of smart annotation collection using crowdsourcing is called *active crowd-labeling.*

Many crowd-labeling problems target to obtain continuous or ordinal labels, such as the position of an object, age of a person, or air temperature. Surprisingly, active crowd-labeling for continuous-valued annotations is a rather sidelined open issue. Related literature on active crowd-labeling mainly focuses on binary annotation problems due to several reasons. First of all, formulating the active crowd-labeling problem in a binary setting is often more tractable with provable mathematical guarantees. Due to the nature of the continuous domain, providing mathematical guarantees in active crowd-labeling solutions proves to be hard, if not impossible. This has pushed the researchers to work with well-studied algorithms by binarizing existing continuous or ordinal annotations. Additionally, presenting the annotation tasks in the form of yes/no or positive/negative reduces task intricacy for the annotators. Although working with binary annotations has several advantages, valuable information is often lost during binarization. Moreover, binary active crowd-labeling approaches are simply impractical when continuous labels are sought. In this work, we estimate the crowd consensus to be used as sample labels from continuous-valued annotations while reducing the cost of the annotation process by employing active crowd-labeling.

We introduce an effective mechanism that decides which sample needs a new annotation and who should annotate it. The method we propose is based on annotator modeling and consensus estimation by Bayesian inference, which is used for producing ordinal and binary labels in addition to continuous labels. One advantage of the method is that it is unsupervised: the gold standard label is not needed for any sample. The proposed method only uses crowd or expert annotations for estimating consensus values and does not depend on the features extracted from the data to be labeled.

In the remainder of this section, we discuss the related work in this domain, followed by the novelty and contributions of this work. In Section 2, we give the details of our proposed active crowd-labeling process. In Section 3, we introduce the datasets on which we evaluate our methods. Section 4 deals with how to use active crowd-labeling to improve existing consensus in crowd-labeling problems. In Section 5, we elaborate on how to conduct smart label collection from scratch and compare our methods with existing methods in the literature. Finally, we present conclusions along with possible future directions in Section 6.

1.1 Crowd-Labeling Literature

Active learning aims to concurrently reduce the training cost and increase the performance of machine learning algorithms by smartly selecting the instances to be included during the learning process. The concept of active learning is a well-suited approach to the crowd-labeling domain where an immense number of annotations need to be acquired, costing both money and time. Settles (2010) surveys and organizes active learning methods, practical considerations, and the relation of active learning to other research areas in detail. Fu, Zhu, and Li (2013) survey the active learning domain from the perspective of instance selection, where active learning methods are categorized into two main groups: those that assume independent and identically distributed instances and those that consider instance correlations. A survey by Frnay and Verleysen (2014) focuses on classification with label noise by defining its sources, and gives a taxonomy on several label noise types.

1.1.1 Active Crowd-Labeling for Binary Annotation Problems

The current literature on active crowd-labeling is mainly focused on binary annotation problems (Sheng, Provost, & Ipeirotis, 2008; Donmez & Carbonell, 2008a, 2008b; Donmez, Carbonell, & Schneider, 2009; Hsueh, Melville, & Sindhwani, 2009; Welinder & Perona, 2010; Yan, Rosales, Fung, & Dy, 2011; Gao, Liu, Ooi, Wang, & Chen, 2013; Lin, Mausam, & Weld, 2016; Tran-Thanh, Venanzi, Rogers, & Jennings, 2013; Tran-Thanh, Huynh, Rosenfeld, Ramchurn, & Jennings, 2014; Fang, Yin, & Tao, 2014; Raykar & Agrawal, 2014; Mozafari, Sarkar, Franklin, Jordan, & Madden, 2014; Nguyen, Wallace, & Lease, 2015; Zhang, Wen, Tian, Gan, & Wang, 2015; Zhuang & Young, 2015; Zhu, Xu, & Yan, 2015; Ho, Jabbari, & Vaughan, 2013; Ho, Slivkins, & Vaughan, 2016; Khetan & Oh, 2016). We briefly survey the main tenets below.

Raykar and Agrawal (2014) model the crowdsourced labeling task sequentially with an epsilon-greedy exploration in a Markov Decision Process. They use a utility function that considers label accuracy, cost and time. Li, Ma, Gao, Su, and Quinn (2016) deal with the budget allocation problem in crowd-labeling by using a Markov Decision Process in a sequential labeling scheme. They propose a trade-off between label quality and quantity. Karger, Oh, and Shah (2011, 2014) define the crowd-labeling problem as a bipartite graph and show results supported by simulated binary data. Their method is inspired by low-rank matrix approximation and belief propagation. Zhuang and Young (2015) verify and investigate the existence of in-batch annotation bias by using a factor graph based batch annotation model on binary data. Ho et al. (2013) formulate the setting as a linear programming problem and work with the dual of the relaxed version. Their method requires the use of gold standard labels for assessing annotator quality and uses weighted majority voting for inferring the consensus. Ho et al. (2016) treat the payment problem for crowdsourcing markets as a multi-armed bandit problem, where each arm represents the contract between a task and an annotator. They propose a method called 'Agnostic Zooming' for selecting the most beneficial contract and study dynamic task pricing. Their work focuses on annotator-sample pairing and deals with binary problems with the task giver's utility function as the main objective.

1.1.2 Sample Selection Strategies

The problem of selecting the most suitable sample has also attracted the interest of researchers. The selection criteria can depend on various factors such as informativeness or uncertainty. Donmez and Carbonell (2008a) study the binary active learning problem by proposing a new sampling strategy. They focus on selecting a suitable sample to include in an unsupervised learning scenario, where the annotator is considered to be infallible. Sheng et al. (2008) use noise-introduced benchmark datasets for sample selection strategies on binary classification problems. Gao et al. (2013) propose an online profit estimation method that weeds out samples which do not need further annotations. Lin et al. (2016) introduce variants of uncertainty sampling and propose impact sampling to select the most informative sample suited for the classifier. Their method decides whether to obtain a new annotation for a readily annotated sample or to introduce a new sample to the crowd-labeled dataset. Khetan and Oh (2016) tackle the problem of binary active crowd-labeling by expending the annotation budget on difficult tasks. They classify high and low confidence tasks in each annotation step and increase the budget allocation for more difficult tasks.

1.1.3 Annotator Selection Strategies

The quality of the annotators varies largely in crowd-labeling problems. Not only do the annotators' expertise vary, but also some of them attempt to exploit the system for profit. Donmez et al. (2009) use the interval estimation learning method for selecting the best annotators by incorporating the exploration-exploitation trade-off. Raykar and Yu (2011) introduce an annotator ranking metric for detecting spammers. Their metric works on binary, categorical, and ordinal labeling tasks. Fang et al. (2014) try to tackle the problem of data scarcity in crowd-labeling by using knowledge transfer from abundant unlabeled data. They report that the approach helps to estimate annotator expertise better and improves performance. Li, Zhao, and Fuxman (2014) propose a crowd targeting framework for selecting the best possible group of annotators for a specific task on binary and categorical data. They introduce information gain as a measure of annotator competence and use EM based top-down and bottom-up approaches for selecting the best annotators. Jagabathula, Subramanian, and Venkataraman (2014) propose a soft penalty scheme for the case of non-malicious annotators for binary labeled data. For each sample, they count the number of times a given annotator agrees with other annotators and calculate the reciprocal of the harmonic mean of such quantities over all samples the given annotator has annotated. A hard penalty scheme is proposed for handling sophisticated adversaries. They use optimal semi-matchings with a quadratic cost function. Zhang et al. (2015) combine a reverse auction model with annotator quality and sample difficulty for conducting crowd-labeling under a budget constraint.

1.1.4 Joint Annotator and Sample Selection

Some of the works in the literature deal with choosing the sample that needs to be annotated along with the most suitable annotator. Donmez and Carbonell (2008b) extend their earlier work (2008a) by considering multiple imperfect annotators and jointly select the optimum annotator-sample pair under a budget constraint. Hsueh et al. (2009) study the annotation selection problem by focusing on annotator noise, class label ambiguity, and the informativeness of a new annotation with regard to the classifier. Tran-Thanh et al. (2013, 2014) investigate the trade-off between budget constraint and annotation quality. Nguyen et al. (2015) use a decision theoretic approach for choosing between acquiring labels from crowds and domain experts. Their method selects a sample and annotator tuple to acquire an annotation. During this process, they account for the active sampling bias and estimate annotator accuracy.

1.1.5 Active Crowd-Labeling for Categorical Annotation Problems

A relatively smaller portion of the existing work in the active crowd-labeling literature concentrates on categorical annotations (Welinder & Perona, 2010; Yan, Rosales, Fung, & Dy, 2011; Mozafari, Sarkar, Franklin, Jordan, & Madden, 2014; Zhu, Xu, & Yan, 2015; Kamar, Hacker, & Horvitz, 2012; Kamar, Kapoor, & Horvitz, 2013, 2015; Venanzi, Guiver, Kohli, & Jennings, 2016). These methods may also be adapted for binary annotations by considering only two categories.

Yan et al. (2011) use uncertainty sampling for sample selection, along with learning annotator expertise on binary and categorical data. Mozafari et al. (2014) propose two active learning algorithms based on sample uncertainty and a classifier's expected error. The methods are tested on a variety of datasets. Zhu et al. (2015) propose an online variant of the Dawid and Skene (1979) algorithm that is motivated by online EM variants and stochastic approximation methods. Kamar et al. (2012, 2013, 2015) use the Galaxy Zoo dataset for the celestial object classification problem. Galaxy Zoo is a crowdsourced effort mainly for the classification of different types of galaxies. Kamar et al. (2012) use Bayesian structure learning to incorporate the human and machine knowledge into the classification task. Kamar et al. (2013) tackle the problem of exploration-exploitation trade-off in worker hiring strategy by modeling the decision-making process as a Markov decision process. Kamar et al. (2015) focus on the problem of rectifying task-related bias of annotators and show that active learning with expert annotators can be used for alleviating bias. Venanzi et al. (2016) use a time-sensitive Bayesian aggregation method to estimate the labeling duration and annotator profile in crowdsourcing systems. They detect bots, spammers or lazy annotators from the duration of their labeling process (either too short or too long). The study is carried out for categorical data.

1.1.6 Passive Crowd-Labeling for Ordinal or Continuous Annotation Problems

Raykar, Yu, Zhao, Valadez, Florin, Bogoni, and Moy (2010) mainly focus on the estimation of consensus by making use of features extracted from the sample data. Their method is also adapted to work without the sample features. The focus of Lakshminarayanan and Teh (2013) is on ordinal labels where task difficulty is incorporated to the discretization of continuous latent variables. Peng, Liu, Ihler, and Berger (2013) propose a domain-specific approach to the protein folding annotation problem by maximizing the log-likelihood of an exponential family mixture model of annotation similarities. Kara, Genc, Aran, and Akarun (2015) deal with the effects of diverse annotator behaviors on consensus estimation for continuous crowd-labeling problems. They also propose a scoring mechanism to determine annotator competence. Ok, Oh, Shin, Jang, and Yi (2017) model the continuous crowd-labeling problem as a bipartite graph and use a belief propagation based Bayesian iterative algorithm when the annotator noise levels are known. For the case where the annotator noise levels are unknown, they employ a non-Bayesian iterative algorithm with marginal performance loss.

1.1.7 Active Crowd-Labeling for Ordinal or Continuous Annotation Problems

Active crowd-labeling for continuous or ordinal valued annotations is a mostly unexplored research area. Marcus, Karger, Madden, Miller, and Oh (2013) make use of gold standard labels to identify low-quality or spammer annotators by a counting approach that combines several binary tasks into an ordinal task. They also identify and avoid coordinated attacks from malicious annotators (i.e. Sybil attacks). Guo, Parameswaran, and Garcia-Molina (2012) deal with the problem of ordering objects in a set by aggregating pairwise comparison of said objects. They devise a maximum likelihood formulation for finding the correct order of objects and show that this problem is NP-hard for their setting where all annotator accuracies are the same. However, their approach to active labeling focuses on the one-shot utilization of the additional budget. Welinder and Perona (2010) tackle the active crowd-labeling problem for continuous-valued annotations, by including the label uncertainty and annotator ability measurement in an EM based approach. Their method detects and excludes spammers during the annotation process and also works on binary and categorical data.

To the best of our knowledge, the approach that we present in this paper and that of Welinder and Perona (2010) are the only approaches which use active crowd-labeling for estimating continuous-valued labels without depending on any prior knowledge about either the annotators (e.g. annotator accuracies) or the samples (e.g. gold standard labels).

1.2 Contributions

Our contributions in this study can be summarized as follows. First, we present one of the few studies on active crowd-labeling for estimating continuous-valued labels from continuous or ordinal valued crowdsourced annotations. We propose two active crowd-labeling methods which produce continuous or ordinal valued consensus labels that can be further converted to binary/categorical labels by quantization, if necessary. The first method, O-CBS, focuses on improving the existing consensuses established from a set of previously collected annotations by selecting a sample-annotator pair for the next annotation. The second active crowd-labeling method, O-CBS+, is an extension of O-CBS. O-CBS+ eliminates the requirement of a readily available annotation set and is able to infer consensuses from scratch by means of annotator exploration/exploitation. Both methods target computational feasibility through a two-tier approach, where choosing a sample with low consensus quality is followed by choosing a high-quality annotator to annotate it. The two-tier approach makes both methods highly scalable and tractable. The proposed methods are data-independent, require no gold standard data to learn annotators, and are specifically designed for problems where the ground truth is not available or easily quantifiable.

Second, based on the variance of the sample's consensus posterior, we provide a novel formulation to estimate sample consensus quality, which corresponds to the total precision of the annotators that annotated the sample. This scoring mechanism prevents budget exhaustion on confusing samples and provides a balanced sample selection.

Third, we address annotator selection problem by introducing a family of annotator competence scoring functions that prevent annotator domination. The dominance suppression mechanism that we introduce prevents ill-intentioned annotators from dominating the system and utilizes high-quality annotators in a balanced manner. We investigate the effects of both sample and annotator selection functions with extensive experiments on nine real-world datasets, two of which are introduced in this paper (Head Pose Annotations Pan and Tilt datasets).

Finally, we study the effects of both a budget induced and a sample consensus quality induced stopping criteria with comparative experiments on all datasets. The results show that O-CBS+ is an effective and budget-friendly (as low as one fifth of the original budget) active crowd-labeling method with high accuracy. Moreover, t-test results prove that it measures up to, or surpasses contender algorithms.

2. Active Crowd-Labeling Methodology

Passive crowd-labeling systems evaluate annotations after the completion of the acquisition phase. Thus, they are easily affected by erroneous annotations given by spammers and inattentive labelers. Each erroneous annotation means money wasted. It is important to be able to distinguish competent labelers from spammers and inattentive labelers early on in the labeling process for acquiring better annotations. Therefore, the most important questions would be: Which sample's label needs to be improved and which annotator should give the annotation? Active crowd-labeling is the process of collecting annotations with such concerns in mind. Smart selection of annotations also result in reduced annotation costs in addition to improved label qualities.

Carrying out a hands-on approach during the annotation acquisition process is in essence similar to active learning from the machine learning domain. In the classical sense, active learning draws its power from selecting the sample to be included in the learning process in a smart manner, thereby producing a well-trained algorithm with fewer samples. In classical active learning, the label of a sample is assumed to be provided by an annotator who always gives correct answers. In contrast, crowd-labeled instances may suffer from low quality annotations. The main motivation behind active crowd-labeling is to simultaneously select the most beneficial annotator-sample pair.

The process of active crowd-labeling is two-fold: One has to make good use of collected annotations, and also make a smart choice about which annotation to request next. We call the first part crowd consensus estimation (Section 2.1). The second stage has two components: how to select the sample to be annotated (Section 2.2) and how to select the annotator to annotate that sample (Section 2.3). Our primary concern is to improve every sample's consensus evenly. Therefore, we select the sample with the lowest consensus quality to be annotated. Once a sample is selected, we select the highest quality annotator for annotating it. This process is repeated with each new annotation in order to even out the sample consensus qualities across the whole dataset.

Algorithm 1 ACL: Active Crowd-Labeling

```
Input:
    Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
function ACL($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}$)
    EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$)
    repeat
        $k \leftarrow$ RequestAnnotation($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, \dots$)
        $\mathcal{K} \leftarrow \mathcal{K} \cup k$          ▷ Add the newly acquired annotation to the annotations set
        EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$)          ▷ Estimate consensus and relearn annotators
    until Budget limit or other stopping criteria are met
end function
```

Our approach consists of iteratively estimating crowd consensus and acquiring new annotations, as outlined in Algorithm 1. In this work, we denote the set of all samples to be annotated, the set of all annotators, and the set of current annotations as $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$, respectively. $\mathcal{J}'$ denotes the annotators that are currently in the system. In Section 2.1, we elaborate on our choice of the EstimateLabels($\cdot$) function used in Algorithm 1, which performs sample consensus estimation and annotator modeling. In Algorithms 2 and 3, we present two different approaches for the RequestAnnotation($\cdot$) function, the details of which are given in Sections 4 and 5, respectively.

2.1 Crowd Consensus Estimation

For crowd consensus estimation, we employ the *Consensus Bias Sensitive Model* (M-CBS) of Kara et al. (2015). The model assumes that a sample i has a single true rate (x_i) and an annotator produces an annotation (y_k) as a function of x_i and their internal decision parameters. In this model, an annotator j is described using four parameters:

- **Adverseness a_j:** The adverseness parameter $a_j \in \{-1, 1\}$ of annotator j describes if the annotator is giving inverted annotations. $a_j = -1$ if the annotator is an adversary and $a_j = +1$ otherwise. a_j has a flat prior.
- **Opinion scale w_j:** Opinion scale $w_j \in \mathbb{R}_{>0}$ of annotator j describes the annotator's tendency to annotate a similar set of samples in a wider/narrower range. $w_j < 1$ means a narrower annotation range, and $w_j > 1$ means a wider annotation range. w_j

has a Gamma prior with hyperparameters selected to assure that it has a mode at 1:

$$w_j \sim \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \tag{1}$$

- **Annotator bias b_j:** Annotators often give positively or negatively biased annotations. Annotator bias parameter $b_j \in \mathbb{R}$ is used for describing this tendency. The prior for b_j in the model is
$$b_j \sim \mathcal{N}\left(b_j; 0, s_B^2\right) \tag{2}$$
- **Precision λ_j:** Precision parameter $\lambda_j \in \mathbb{R}_{>0}$ describes the annotator's consistency while providing annotations. Its prior is
$$\lambda_j \sim \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \tag{3}$$

Using these parameters and the true rate x_i, the model describes the annotation y_k as a random variable with the probability distribution

$$y_k \sim \mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \tag{4}$$

where i_k and j_k are the sample and annotator of the annotation k, respectively.

Given the values y_k, the aim is to estimate consensuses on the true rates (x_i) of samples while simultaneously estimating the annotator parameters using maximum a posteriori estimation.

2.2 Which Sample Needs a New Label?

Since we want to improve our consensus estimations for the samples, we are in need of getting more annotations. Instead of randomly selecting samples for requesting annotations, a smarter strategy would reduce annotation costs while attaining high quality consensuses. The process of choosing which sample to annotate in a timely manner is of utmost importance since active crowd-labeling is a real-time process. Calculating the utility of all possible sample-annotator pairings for finding the optimal solution is often computationally very complex (at least $\mathcal{O}(nm)$) and poses scalability problems for large datasets and open annotator marketplaces. To this end, we opt for adopting a sub-optimal yet still beneficial approach to predict samples with low consensus quality by making use of readily available parameters inferred during the active crowd-labeling process.

During active crowd-labeling, our knowledge of a sample's consensus is gathered in its posterior distribution. Our motivation comes from the observation that a sample's quality may roughly be assessed by the variance of this posterior distribution. Since the system state changes in every annotation step (the addition of a new annotation), we use the superscript (t) for referring to the system parameters at the annotation step t. Using Bayesian rule on the full joint probability of the M-CBS model, we find the posterior distribution of the consensus $x_i^{(t)}$ at the annotation step t as

$$x_i^{(t)} \Big| \left\{ y_k, \theta_{j_k}^{(t)} : k \in \mathcal{K}_i^{(t)} \right\} \sim \mathcal{N}\left(x_i^{(t)}; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k}^{(t)} \left({w_{j_k}^{(t)}}^{-1} a_{j_k}^{(t)} y_k - b_{j_k}^{(t)} \right)}{\sum\limits_{k:i_k=i} \lambda_{j_k}^{(t)}}, \left(\sum_{k:i_k=i} \lambda_{j_k}^{(t)} \right)^{-1} \right) \tag{5}$$

where $\theta_j^{(t)} = \left\{a_j^{(t)}, w_j^{(t)}, b_j^{(t)}, \lambda_j^{(t)}\right\}$ is the set of parameters of annotator j inferred at annotation step t and $\mathcal{K}_i^{(t)} = \{k \in \mathcal{K} : i_k = i\}$ is the set of annotations of sample i. The derivation of this distribution is provided in Appendix A.

The smaller the variance of this distribution, the more confident we are on the inferred consensus and we want to request new annotations for the samples that we are less confident about. Thus, we use the reciprocal of the variance as a measure of consensus quality, namely the consensus quality score $S_S(i)$ of sample i

$$S_S(i) = \sum_{k:i_k=i} \lambda_{j_k}^{(t)} \tag{6}$$

where λ_{j_k} are the precision parameters of every annotator j that has annotated sample i. This is equivalent to counting the annotations of a sample weighted by its annotators' precision. Thus, the consensus quality of a sample is only as good as the annotators' precision that have annotated it. Additionally, it also ensures that a sample's annotation count is also incorporated into its quality assessment. Note that adding a new annotation to an existing sample will definitely increase the sum and decrease the variance since λ values are positive. From a budget minimization point of view, it would be more beneficial to concentrate on those samples with the lowest scores. The approach that we present here is a fast (with complexity $\mathcal{O}(n)$) and reasonable way to reduce annotation costs and improve on the consensus values.

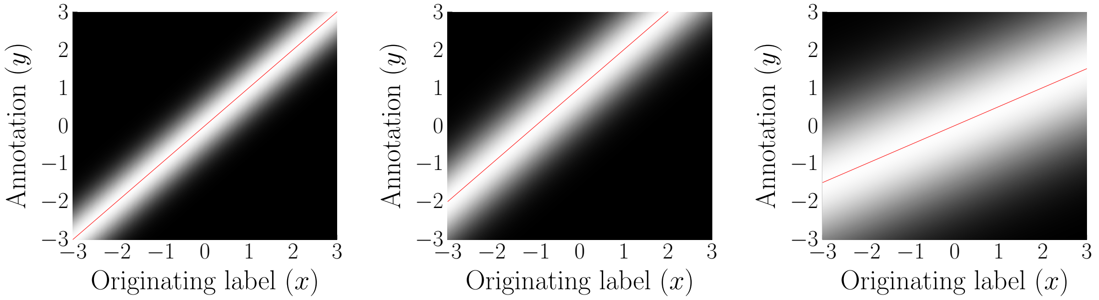
2.3 Who Annotates Better?

During the active crowd-labeling process, we need to identify competent annotators to utilize for new annotations. Thus, we need to rate annotators based on their competences. As the annotator competence scoring mechanism, we refer to the formulation for M-CBS (Kara et al., 2015). The score is described as "the sum of the joint probabilities of all possible annotations that can be produced by an annotator and the most probable originating label for those annotations given the annotator parameters" (Kara et al., 2015). The formulation for the annotator score is

$$S_A(j) = \frac{1}{{w_j^{(t)}}^2} \sqrt{\frac{\lambda_j^{(t)} \left(1 + {w_j^{(t)}}^2\right)}{2\pi}} (e_j - d_j) \tag{7}$$

where $d_j = \min\{c, \max\{w_j^{(t)}(b_j - c), -c\}\}$, $e_j = \max\{-c, \min\{w_j^{(t)}(b_j^{(t)} + c), c\}\}$, and $[-c, c]$ defines the annotation range. The score is derived by calculating the path integral of $p(x, y|\theta)$ along the linear mapping that defines the annotator, where d_j and e_j are the upper and lower limits of the path integral (for derivation details, see Kara et al., 2015).

This formulation ensures that the annotator competence score is high when w_j is close to 1 and b_j is close to 0, which are desirable for the annotators to produce annotation values close to the true rate. Additionally, the annotator competence score also increases with higher λ_j to select more consistent annotators. In Figure 1, we present three examples of annotators commonly encountered in crowd-labeling problems.



(a) Very competent

(b) Positively biased

(c) Inattentive

Figure 1: Three examples of annotators: Very competent, positively biased, and inattentive. Grayscale values represent posterior probability of annotation value ($p(y|x, \theta)$); the higher the intensity, the higher the probability. The red line is the peak of this distribution. For very competent annotators, w_j is close to 1 and b_j is close to 0. Additionally, they have high λ_j values resulting in a concentrated band of annotations around the peak. In contrast, inattentive annotators have lower λ_j values which result in more scattered annotations.

3. Crowd-Labeled Datasets Used for Evaluation

We evaluate the results of the proposed active crowd-labeling method using nine real datasets: two Head Pose Annotations datasets (tilt, pan) which are introduced in this paper, the Kara Age Annotations dataset (Kara et al., 2015) and six Affective Text Analysis datasets (anger, disgust, fear, joy, sadness, surprise) of Snow, O'Connor, Jurafsky, and Ng (2008). Table 1 summarizes the datasets used in this work.

Dataset	Annotations	Samples	Annotators	Ground Truth Range	Annotation Range
Head Pose Annotations: tilt, pan (introduced in this work)	5399	555	189	$\{-90, \ldots, 90\}$	$\{1, \ldots, 7\}$
Kara Age Annotations (Kara et al., 2015)	10020	1002	619	$\{0, \ldots, 69\}$	$\{1, \ldots, 7\}$
Affective Text Analysis: anger, disgust, fear, joy, sadness, surprise (Snow et al., 2008)	1000	100	38	$\{0, \ldots, 100\}$	$\{0, \ldots, 100\}$

Table 1: Annotation datasets used in this work. For evaluating our work, we introduce head pose annotations dataset including tilt and pan modalities. Additionally, we use Age Annotations dataset of Kara et al. (2015) and six Affective Text Analysis datasets of Snow et al. (2008). For all datasets, the ground truth values and the annotations are in the continuous domain.

3.1 Head Pose Annotations Datasets

In this paper, we introduce the Head Pose Annotations datasets for our evaluations. For the annotation tasks, we used head pose images obtained from the Head Pose Image Database (Gourier, Hall, & Crowley, 2004). This database contains head poses of 15 people, with different head orientations (i.e. tilt and pan combinations.) The pan values range from -90 to 90 degrees with 15 degree increments. The tilt values are in the set {-90, -60, -30, -15, 0, +15, +30, +60, +90}. Although there are 117 possible tilt and pan combinations, the Head Pose Image Database omits head poses for extreme tilt cases and contains 93 different head orientations. Out of these 93 head orientations, we chose the photos having tilt and pan values of {-90, -60, -30, 0, +30, +60, +90} degrees for the annotation tasks, due to budgetary reasons. Figure 2 shows a single subject's all possible head pose combinations that we chose for getting annotated. During the annotation tasks, each image sample is annotated for both tilt and pan modalities, thus resulting in two separate datasets. Each of these pan and tilt datasets consist of 5399 annotations attributed to 37 distinct head poses from 15 subjects, making up a total of 555 head pose images.

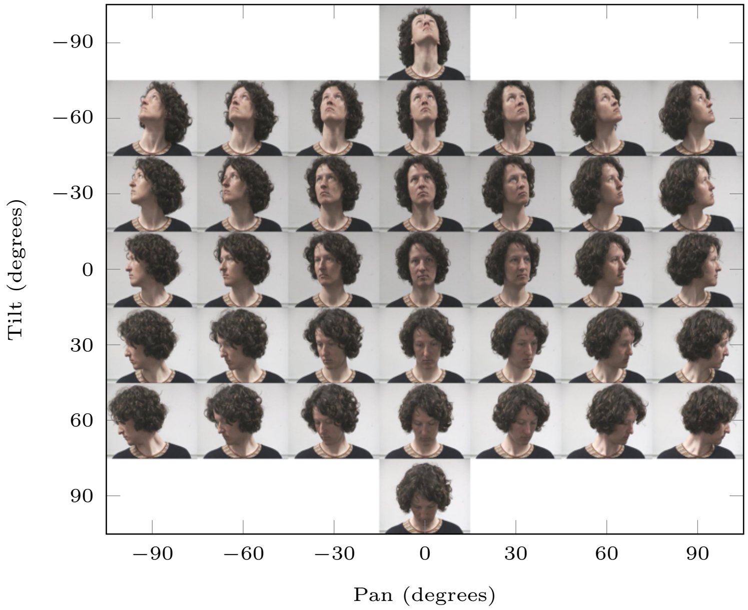
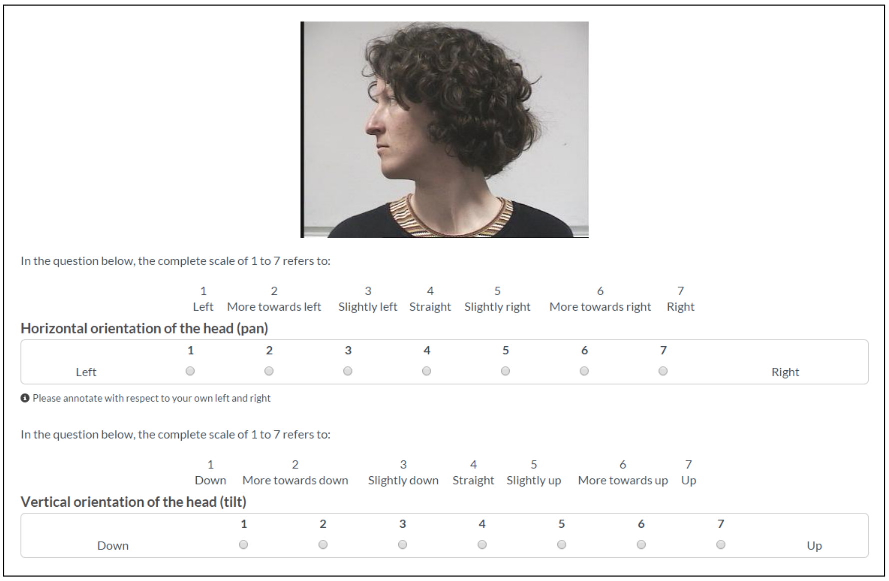


Figure 2: 37 distinct head poses of a person, which are chosen for the annotation tasks in the Head Pose Annotations datasets. The head pose images are taken from the Head Pose Image Database (Gourier, Hall, & Crowley, 2004).

The annotations were collected using CrowdFlower. In the CrowdFlower platform, we prepared a questionnaire in which the annotators are shown a head pose photo and asked about the head orientation. For each photo, we asked the participants to annotate: (a) the horizontal orientation (pan) according to their own left and right in the range 1 (left) to 7 (right), (b) the vertical orientation (tilt) in the range 1 (up) to 7 (down). Figure 3 shows a sample of what the annotators see when they are working on our head orientation annotation task. In each questionnaire, the annotators were asked to annotate a batch of



In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Left	More towards left	Slightly left	Straight	Slightly right	More towards right	Right

Horizontal orientation of the head (pan)

	1	2	3	4	5	6	7	
Left	○	○	○	○	○	○	○	Right

Please annotate with respect to your own left and right

In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Down	More towards down	Slightly down	Straight	Slightly up	More towards up	Up

Vertical orientation of the head (tilt)

	1	2	3	4	5	6	7	
Down	○	○	○	○	○	○	○	Up

Figure 3: Sample questionnaire for the head pose annotation process.

10 head pose photos. The annotators were free to annotate more than one batch or to leave the system early and provide an incomplete batch. In order to assist the annotation process, we also provided the annotators with verbal descriptions of each possible ordinal assessment. The annotators were asked to provide pan annotations with respect to their own left and right to avoid unnecessary confusion. The order of placement for the possible answers also targeted to avoid confusion, where the answer "left" appeared on the left-hand side of the questionnaire and vice versa.

Sample annotation count	**7**	**8**	**9**	**15**	**16**	**17**
Number of samples	10	10	475	6	34	20

(a) Number of annotations per sample for the Head Pose Annotations Datasets

Annotator workload	**5**	**10**	**17**	**20**	**24**	**30**	**39**	**40**	**45**	**50**	**55**	**60**	**70**	**75**	**80**	**84**	**90**	**100**
Number of annotators	1	61	1	45	1	26	1	15	2	13	1	7	5	1	4	1	2	2

(b) Annotator workloads for the Head Pose Annotations Datasets (the number of annotations made by an annotator)

Table 2: Statistics for the Head Pose Annotations Datasets

In Table 2a, we present the annotation frequency of the samples. Out of 555 samples, 475 have 9 annotations, with other samples having as few as 7 and as many as 17 annotations. Table 2b shows the annotation frequency of the annotators, which we call annotator workload. A total of 189 annotators participated in the annotation tasks. Most common annotator workloads are multiples of 10 since many annotators completed the batch tasks assigned to them. For example, 61 annotators annotated 10 samples and 2 annotators annotated 100 samples.

3.2 Kara Age Annotations Dataset

The Kara Age Annotations dataset (Kara et al., 2015) consists of a total of 10020 annotations of 619 annotators on 1002 samples. CrowdFlower was used for annotating the samples of the FGNet Aging Database. In the FGNet Aging Database, each sample is a picture of a person with known age between 0 and 69. The annotators were asked to rate the age of the person in the range of 1 to 7 where smaller rate means younger. 10 annotations were collected for each sample.

3.3 Affective Text Analysis Datasets

We conduct our third set of tests on the six Affective Text Analysis datasets. Each of these datasets has 1000 annotations on 100 short news headlines, drawn from various news sources (Strapparava & Mihalcea, 2007), regarding positive and negative emotions. The task is to annotate a headline for each emotion, namely anger, disgust, fear, joy, sadness, and surprise. The annotators were asked to provide annotations in the interval of 0 to 100 for each emotion. 10 annotations per task were collected from 38 annotators using Amazon Mechanical Turk. The provided ground truth values are the averages of expert opinions.

Annotating emotions is a highly subjective task. There is no quantitative metric with which to measure the intensity of an emotion. Thus, the best possible approach is to consult experts and accept combinations of their opinions as the ground truth labels. However, comparing estimated labels obtained from crowd annotations with these ground truth values only establishes how well the crowd can estimate the average opinion of experts. Thus, it is very likely that high quality crowd opinions may be dismissed as subpar since they differ from the ground truth produced by only a few experts.

It is more common to express one's emotions in a state of existent/non-existent instead of on a scale of 0 to 100. Similarly, it is not easy for the annotator to annotate the emotion on such a fine scale. Therefore, a more practical approach is to compare the crowd's opinions against the experts' after binarization.

In light of these issues, we compare the binarized estimated labels with the binarized ground truth values for the six Affective Text Analysis datasets, as has been done in previous works that use this data (Raykar & Agrawal, 2014). Although we binarize the estimated output labels, we use the input annotations from the crowd as they are. By not binarizing the input annotations, we prevent the loss of valuable information, which may prove crucial for borderline decisions. Therefore, the results for Affective Text Analysis datasets are given as accuracies.

For all nine datasets, annotations are linearly mapped to the range $[-3, 3]$ before processing. This is done to preserve compatibility with the hyperparameters chosen by Kara et al. (2015). The results for the Head Pose Annotations datasets and the Kara Age Annotations dataset are given in mean absolute degree and age error, respectively. Therefore, their inference results, which are in the range $[-3, 3]$, are linearly mapped to their related ground truth ranges (i.e. $[-90, 90]$ degrees and ages 0 through 69.)

As we mention in Section 1.1, the work of Welinder and Perona (2010) is the only approach besides this work that estimates continuous valued labels by means of active crowd-labeling without depending on any prior knowledge about either the annotators or the samples. Thus, on the Head Pose Annotations and the Kara Age Annotations datasets, we compare our results with the work of Welinder and Perona. We also provide binarized comparisons with the work of Raykar and Agrawal (2014) on the six Affective Text Analysis datasets.

4. O-CBS: Improving the Existing Consensus Using Active Crowd-Labeling

When dealing with annotation problems, the task at hand often requires working with a limited pool of annotators, especially when the subject requires expert annotators. However, due to budget and/or time constraints, each annotator annotates only a subset of all samples. Although we can infer a preliminary consensus, later on we may want to reconsult the same annotators for the samples that they did not annotate beforehand in order to improve the consensuses.

In this section, we propose an annotation collection and consensus improvement method for the situation mentioned above, which we call O-CBS (Online M-CBS). Algorithm 2 gives the details of the annotation requesting mechanism for improving the existing consensus. We first need to identify which sample's consensus is not satisfactory and needs to be improved the most. The algorithm expects a sample consensus quality scoring function which measures trustworthiness of the consensus estimation and gives higher results when the estimation on the consensus is more trustworthy. Then, the sample with the least consensus quality score is selected to be improved. The sample consensus quality score function introduced in Equation 6 is a suitable choice.

Algorithm 2 RequestAnnotation: Requesting annotation for improving the existing consensus

Input:

Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
i_k and j_k are the sample and annotator of annotation k, respectively
$S_S(\cdot)$ and $S_A(\cdot)$ are the sample consensus quality function and annotator competence scoring function, respectively. (We assume that S_S and S_A are intrinsically aware of the annotator parameters a, w, b, and λ)

Output: New annotation k

```
function RequestAnnotation($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, S_S(\cdot), S_A(\cdot)$)
    for all $i \in \mathcal{I}$ do
        $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$        ▷ Annotations of sample $i$
        $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$        ▷ Annotators of sample $i$
    end for
    $i \leftarrow \underset{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset}{\arg\min} S_S(i')$        ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample
    $j \leftarrow \underset{j' \in \mathcal{J}' \setminus \mathcal{J}_i}{\arg\max} S_A(j')$        ▷ Select the most competent annotator from the set of active annotators who had not annotated sample $i$
    $k \leftarrow$ Request an annotation for sample $i$ from annotator $j$
    return $k$
end function
```

The second part of the problem is the selection of the most suitable annotator for the selected sample. For this, we need an annotator competence scoring function that gives higher scores for more competent annotators. Finally, we ask the annotator with the highest competence score for a new annotation for the selected sample.

O-CBS is based on Algorithm 1 with M-CBS as the EstimateLabels($\cdot$) function and Algorithm 2 as the RequestAnnotation($\cdot$) function. In this setting, RequestAnnotation($\cdot$) employs S_S (Equation 6) as the sample consensus quality scoring function. We investigate a family of annotator competence scoring functions, and we denote O-CBS with such different functions $(S_A, S_A^{\mathcal{K}}, S_A^1, \dots)$ as O-CBS($\cdot$). As a baseline method, we use O-CBS$\big(S_A^{\mathcal{R}}\big)$ which employs S_S for sample selection but selects annotators randomly. As another baseline method, we use O-CBS(*Random*) which is a special case where the sample consensus quality scoring and the annotator competence scoring functions are both replaced with random selection.

4.1 Effectiveness of the Sample Scoring Function S_S

Since S_S is our choice of sample selection strategy in O-CBS, we start with presenting its performance by comparing it against random sample selection. In Figure 4, we observe the effectiveness of using the sample scoring function S_S across all nine datasets. We report the MAE on the Kara Age Annotations and the Head Pose Annotations datasets. On the Affective Text Analysis datasets we report the accuracy. The graphs show that S_S is a favorable sample selection strategy across all datasets in terms of mean absolute error and accuracy. Especially in pan, anger, joy, and sadness datasets, there is a significant improvement over random sample selection. Although O-CBS$\big(S_A^{\mathcal{R}}\big)$ falls behind O-CBS(*Random*) in the *fear* and *surprise* datasets as the number of annotations increases, the overall performance of S_S is beneficial. Even in the absence of an annotator selection strategy, S_S by itself provides significant improvement to active crowd-labeling performance.

4.2 Balancing the Scales: Suppressing Annotator Domination

The annotator competence scoring function described in Equation 7 satisfies the aforementioned requirement of giving higher scores for more competent annotators. In this section, we discuss the shortcomings of the said annotator competence scoring function and propose several updates to alleviate these shortcomings.

Since our focus is on crowd annotation problems without any gold standard, we trust the consensus of the crowd to be true. However, it is possible that the majority of the crowd might be wrong or ill-intentioned. Moreover, ill-intentioned annotators are inclined to annotate more samples for gaining more money, resulting in an unbalanced system.

The stability of a crowd grows when more people are in it and the crowd-labeling approach is more susceptible to the actions of said people when the crowd is small. If the system is dominated by incompetent annotators, whenever a competent annotator joins the system, their opinion will be treated as an outlier and good annotators will have a low annotator competence score due to the mechanism introduced in Section 2.3. Since the active crowd-labeling method is inclined to acquire new annotations from the high scoring annotators, the method will continue requesting annotations mainly from incompetent annotators. Even if more truly competent annotators join the system, it may prove to be challenging to balance the scales in favor of them. Therefore, it is crucial to prevent annotator overloading early on and to let the method concentrate on competent annotators later on.

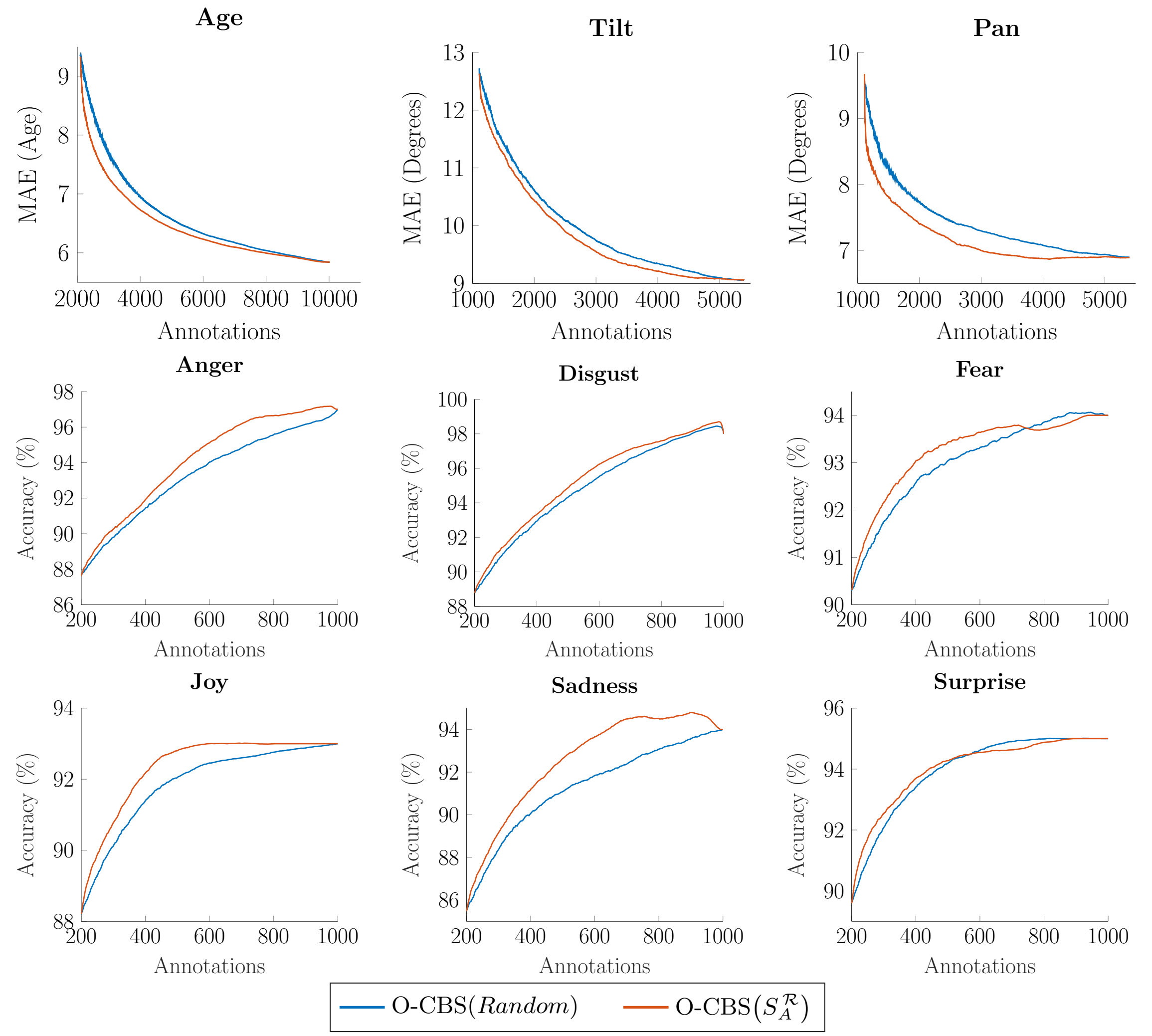


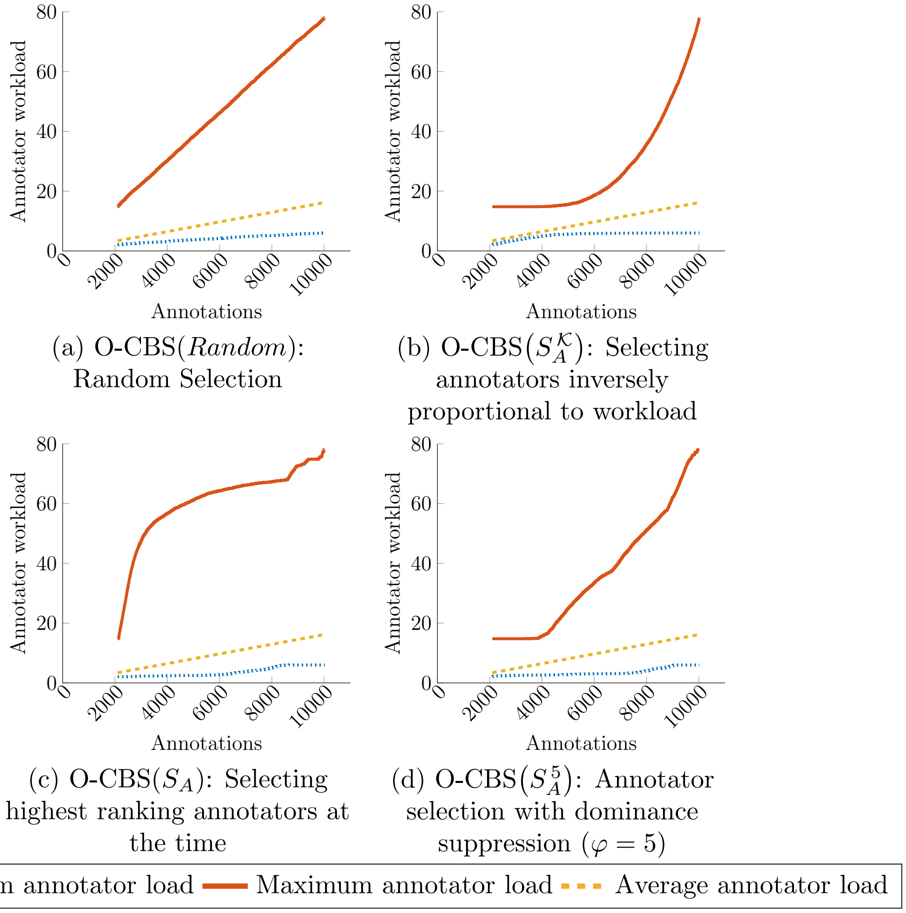
Figure 4: Effect of using S_S for sample selection on the Kara Age Annotations, the Head Pose Annotations, and the Affective Text Analysis datasets, averaged over 100 runs with different starting subsets. On the top row of figures we report the MAE and on the other rows we report the accuracy. O-CBS($Random$) employs both random sample and random annotator selection, whereas O-CBS$\left(S_A^{\mathcal{R}}\right)$ employs random selection only for annotators and uses S_S for sample selection.

For overcoming these issues, we introduce a weighting factor to the annotator scoring mechanism proposed in Section 2.3. The idea is to suppress the annotator scores $S_A(j)$ proportionally to the annotator workloads so that the score of highly loaded annotators are suppressed. Additionally, we want to reduce this effect as the system gets more reliable in terms of annotations. We call this weighting factor *the dominance suppression factor*, which is

$$\left|\mathcal{K}^j\right|^{-\varphi \frac{\left|\mathcal{J}^1\right|}{|\mathcal{K}|}} \tag{8}$$

where $\varphi > 0$ is *the dominance suppression coefficient* which controls the effect of the weight, $|\mathcal{K}|$ is the current number of annotations, $|\mathcal{K}^j|$ is the number of annotations of annotator j, and $|\mathcal{J}^1|$ is the number of annotators that have at least one annotation.

$\frac{|\mathcal{K}|}{|\mathcal{J}^1|}$ is the average number of annotations per annotator. With each new annotation, this factor increases; with each new annotator, it decreases momentarily. New annotator introduction to the system is rarer than adding new annotations to the annotation pool from current annotators. Thus, the suppression effect of the newly introduced dominance suppression factor almost always decreases as the active crowd-labeling process progresses.



(a) O-CBS($Random$): Random Selection

(b) O-CBS$\left(S_A^{\mathcal{K}}\right)$: Selecting annotators inversely proportional to workload

(c) O-CBS(S_A): Selecting highest ranking annotators at the time

(d) O-CBS$\left(S_A^5\right)$: Annotator selection with dominance suppression ($\varphi = 5$)

Figure 5: Change in the minimum, maximum, and average annotator workloads during the active crowd-labeling process. The results are provided for the Kara Age Annotations dataset. The horizontal axis represents the total number of annotations currently in the system. The vertical axis represents the number of annotations (workload) of the annotator in question. Note that each point on the plots may represent a different annotator. Depending on the annotator selection criterion, the maximally and minimally loaded annotators will change during the annotation process.

Thus, we introduce a dominance suppression based annotator competence score as the product of the annotator competence score (Equation 7) and the dominance suppression factor (Equation 8):

$$S_A^{\varphi}(j) = S_A(j)\left|\mathcal{K}^j\right|^{-\varphi\frac{\left|\mathcal{J}^1\right|}{|\mathcal{K}|}} . \tag{9}$$

As a baseline method, we also introduce a simple annotator score based only on the annotator's workload:

$$S_A^{\mathcal{K}}(j) = |\mathcal{K}^j|^{-1}. \tag{10}$$

Figure 5 shows the load of minimum, maximum and averagely loaded annotators. In Figure 5a where new annotations are randomly selected, maximum annotator load increases linearly and diverges quickly from the average load. This means that only a handful of annotators are dominating the system. This is a tendency that we aim to avoid as mentioned before.

If $S_A^{\mathcal{K}}(j)$ is used as the annotator score, we see that the maximum annotator load tends to stay the same for a long time (Figure 5b). Although this behavior is desired since it prevents domination by a group of annotators, this scoring mechanism by its very nature does not incorporate the behavior of the annotator and fails to pinpoint competent annotators.

When the scoring function $S_A(j)$ (Section 2.3) is used, the active crowd-labeling system tends to overload the high scoring annotators and the maximum load increases rapidly (Figure 5c). However, this is risky due to the problems described earlier.

When dominance suppression is active, the scores of highly loaded annotators are weighted down for obtaining the desired behavior. In Figure 5d, we choose the dominance suppression coefficient $\varphi = 5$ and it is clear that we reach a more stable annotator load distribution. Early on in the active crowd-labeling process, the maximum annotator load holds steady while the system gets acquainted with the annotators in an objective manner. After a while the maximum workload starts to increase with the diminishing effect of the dominance suppression factor, thereby utilizing high quality annotators.

4.3 Effects of Annotator Dominance Suppression

In this section, we will discuss the results of improving the existing consensus by using active crowd-labeling under several different dominance suppression criteria. However, the data described in Section 3 was not collected considering active crowd-labeling. Thus, first we need to create starting subsets of the annotation data for evaluating O-CBS. We present our results on nine datasets, namely the Kara Age Annotations, the Head Pose Annotations, and the six Affective Text Analysis datasets.

4.3.1 Selecting Starting Subset for Active Crowd-Labeling:

Assume that annotations are already collected for a fixed sample set and we want to improve the consensus values without adding new annotators to the system. This is a common case in many institutions where a dataset is collected and annotated in-house. In this setting, the problem of extending the annotation dataset boils down to asking an annotator to annotate a sample that they have not annotated before. In order to emulate this, we create annotation subsets for each dataset that satisfy the following conditions:

- Every sample has an annotation
- Every annotator has at least 2 annotations
- Every annotator has an annotation for a sample that also has an annotation from another annotator (this is needed for being able to compare annotators)

For each dataset, we prepare 100 different subsets satisfying these conditions. We fix the subset sizes, i.e. number of annotations, to 2100 for the Kara Age Annotations dataset, 1110 for the Head Pose Annotations datasets, and 200 for the Affective Text Analysis datasets. The details of the subset selection algorithm are given in Appendix B. In Table 3, we give pairwise inter-set similarity statistics of the created subsets. We observe that there is approximately 20% overlap between the resulting subsets on average. This similarity is low enough to ensure that the results of our active crowd-labeling scheme do not depend on initial conditions.

Dataset	Subset size	Inter-Set Similarity (%)		
		Min	Average	Max
Head Pose Annotations	1110	15.68	21.47 ± 1.17	26.13
Kara Age Annotations	2100	18.43	21.13 ± 0.85	24.81
Affective Text Analysis	200	11	19.95 ± 2.63	29.5

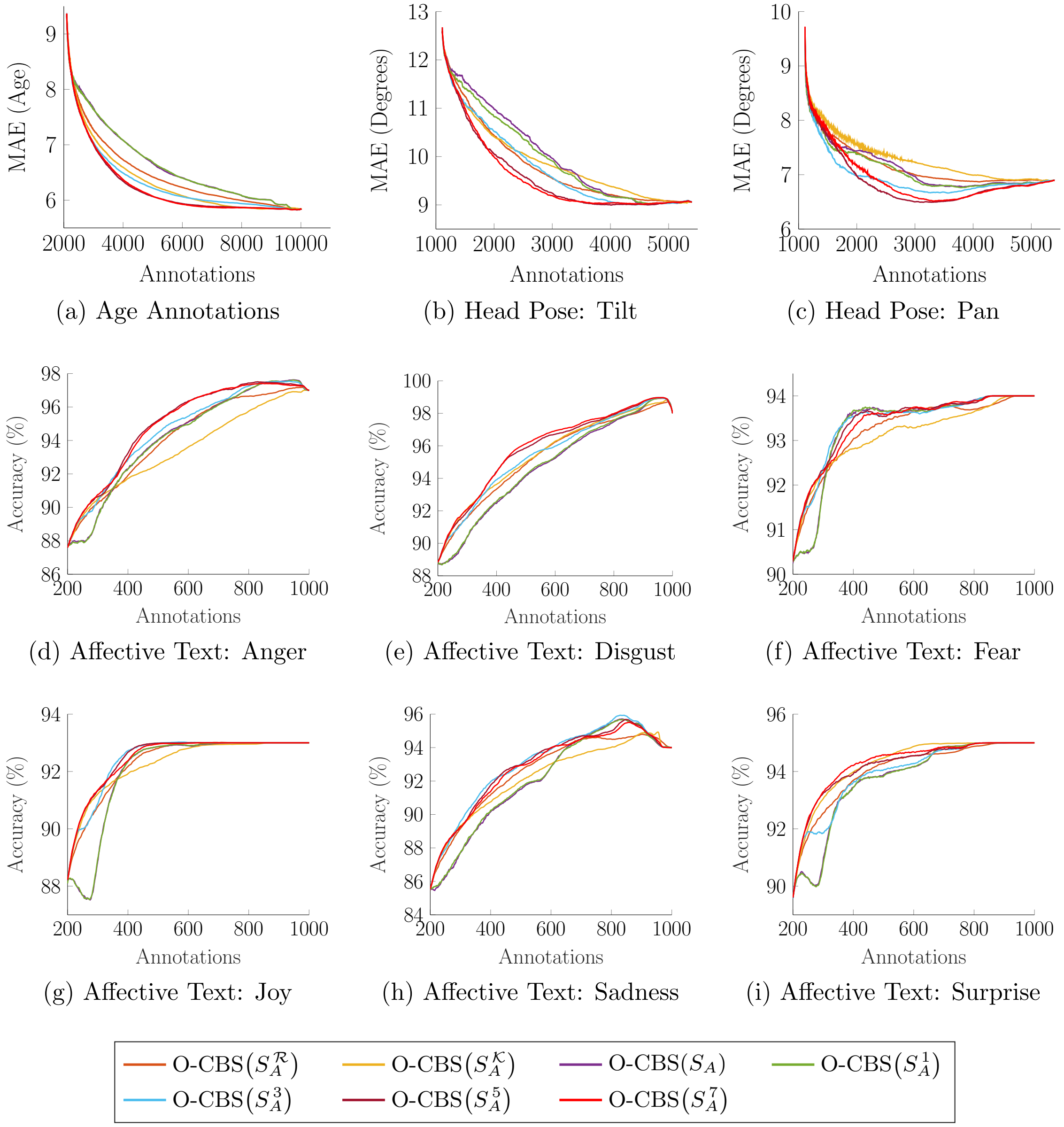
Table 3: Details of the created subsets

4.3.2 Mean Absolute Age Error Improvement on the Kara Age Annotations Dataset:

In Figure 6a, we present the results of our method's effect on mean absolute error in terms of age by trying out different dominance suppression coefficients φ on the Kara Age Annotations dataset. We have two baseline methods that we compare our approach with. The first is O-CBS$\big(S_A^{\mathcal{R}}\big)$ where the annotator is selected randomly. The second is where the sample with the worst consensus quality score is annotated by the annotator with the least annotation count (O-CBS$\big(S_A^{\mathcal{K}}\big)$). We do not plot O-CBS(*Random*) curves in Figure 6, since we already gave their comparison with O-CBS$\big(S_A^{\mathcal{R}}\big)$ in Figure 4. When φ is small, our method fails to suppress low-quality annotators as we describe in Section 4.2, resulting in even lower performance than the baseline methods. When $\varphi \geq 3$ our method outperforms the baseline approaches significantly. Instead of collecting 10000 annotations, roughly 6000 annotations are sufficient to drop below 6 years in terms of mean absolute error.

4.3.3 Mean Absolute Degree Error Improvement on the Head Pose Annotations Dataset:

We further test the performance of O-CBS on the Head Pose Annotations *tilt* and *pan* datasets. Figures 6b and 6c show the change in the mean absolute error in degrees, according to different dominance suppression coefficients. Similar to the performance on the Kara Age Annotations dataset, O-CBS performs subpar when the dominance suppression coefficient φ is small, or the non-suppressed annotator scoring mechanism $S_A(j)$ is used. On the *tilt*



(a) Age Annotations

(b) Head Pose: Tilt

(c) Head Pose: Pan

(d) Affective Text: Anger

(e) Affective Text: Disgust

(f) Affective Text: Fear

(g) Affective Text: Joy

(h) Affective Text: Sadness

(i) Affective Text: Surprise

Figure 6: Improving the existing consensus on the Kara Age Annotations, the Head Pose Annotations, and the Affective Text Analysis datasets. For the Head Pose Annotations and the Kara Age Annotations datasets, we report the mean absolute error (MAE) with different annotator competence scoring functions. For the Affective Text Analysis datasets, we report the accuracy. The curves are averaged over 100 runs with different starting subsets.

dataset, the MAE achieved at the end of the annotation procedure can be achieved earlier on with much fewer annotations by using $\varphi \geq 5$. For the *pan* dataset, we also observe that the curves with $\varphi \geq 5$ have a trough shape around 3000 annotations. This trend is due to the fact that high-quality annotators are distinguished early on, resulting in low error.

Additional annotations provided by lower quality annotators result in degrading the system performance. Note that we let the system to use all annotations for examining the total effect of the annotations on consensus quality. Every point on these graphs actually show the performance at the corresponding annotation limit. Therefore, it is also possible to interpret Figure 6 as what the performance of the system will be, should a budget limit be enforced.

4.3.4 Accuracy Improvement on the Affective Text Analysis Datasets:

We also test our method on the six Affective Text Analysis datasets, which present a more challenging problem since the datasets are much smaller than both the Kara Age Annotations and Head Pose Annotations datasets. Our first observation in Figures 6d to 6i is that each dataset belonging to an emotion results in different baseline method characteristics, presenting diverse conditions in which we test our method.

In consort with the results in Figures 6a to 6c, a higher dominance suppression coefficient $\varphi \geq 5$ helps to achieve high accuracy with fewer annotations. This effect is most prominent in *fear*, *joy*, and *surprise* datasets where roughly 400 out of 1000 annotations are sufficient for achieving near-maximum accuracy. Additionally, introducing the dominance suppression factor helps us to outperform the two baseline methods significantly, specifically in the *anger*, *fear*, and *sadness* datasets.

4.4 Speeding Up the Inference Process

In passive crowd consensus estimation, we randomly initialize the annotator parameters and iteratively infer the resulting annotator parameters using the model described in Section 2.1. In an active crowd-labeling process, this inference process is repeated with each new annotation and the computational cost increases duly. However, we expect a small change in annotator parameters since there is only a small change in the annotations set. Thus, we can use our previous knowledge about the annotator parameters to reduce the complexity of the process.

M-CBS describes an annotator using a linear map and a noise parameter. When there are only a few annotations of an annotator, the model might infer a wrong conclusion about the behavior of the annotator in question. This is a very common case especially in the early phases of the active crowd-labeling scheme.

In Figure 7, we present three random initialization approaches and their effect on iteration count and MAE. The first approach is to initialize every annotator's parameters each time a new annotation is acquired, thus avoiding sticking to a local extremum. This is actually a baseline approach which results in high iteration counts, especially early on in the active crowd-labeling process. Alternatively, we may initialize the parameters of every annotator that has provided an annotation for the newly annotated sample, since the new annotation will affect the sample's consensus. It is also possible to take a more conservative approach and reinitialize the parameters of only the new annotation's annotator. Both of these approaches still have the advantage of avoiding being stuck at local extrema. Results show that both of these approaches result in a significantly decreased number of iterations, with the latter approach being lower in iteration numbers. There is no change in the MAE,

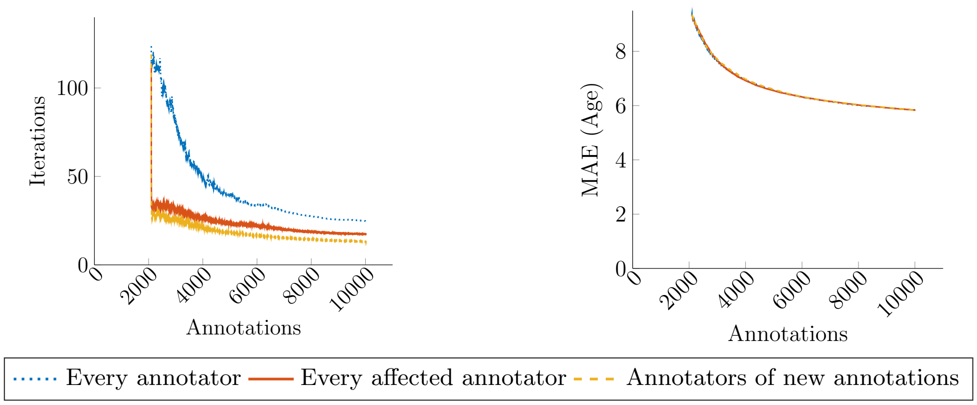


Figure 7: The effect of three different random initialization approaches on the number of iterations for O-CBS(*Random*) (random annotation addition). The results are provided for the Kara Age Annotations dataset. Reinitializing the annotator parameters of only those providing new annotations results in much fewer iterations with the same MAE.

which confirms that these time-saving methods do not affect the quality of the consensus estimation process.

5. O-CBS+: Starting Active Crowd-Labeling from Scratch

When the task giver has full control over the label collection process, it is more beneficial to identify the annotator quality as soon as possible. Timely evaluation of annotator quality results in saving both money and time by achieving high quality consensuses using fewer annotations. Thus, it is important to use the active crowd-labeling process from scratch.

O-CBS handles the case when we are already acquainted with the annotators, thus have an opinion about their annotation behaviors. However, for using active crowd-labeling at the start of the crowd-labeling process, we need to not only utilize current annotators, but also assess new annotators.

Even though the sample pool is fixed at the end, every sample seems to be new at the early stages of active crowd-labeling since we do not have annotations for them. O-CBS is not designed for the addition of new samples. When a new sample needs to be annotated, it is crucial to have an opinion about its consensus in a timely fashion.

In Algorithm 3, we take these concerns into account. We first check whether there is a new sample or not. If there are new samples that have not been annotated before, we randomly select a sample to be annotated. Otherwise, we select the sample with the worst consensus quality score, similar to O-CBS. Upon the selection of the sample, we need to decide if we want to have this sample annotated by a known annotator (exploit) or a new annotator (explore). If we decide to exploit an annotator, we request an annotation for the selected sample from the highest scoring available annotator. When exploring a new annotator, we want to have at least two annotations of the annotator since we want to have an opinion about their behavior and one of the annotations should be of an already annotated sample. Thus, we request two annotations from the new annotator accordingly.

Algorithm 3 RequestAnnotationExp: Requesting annotation for smart label collection from scratch

Input:

Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
i_k and j_k are the sample and the annotator of the annotation k, respectively
$S_S(i)$ is the consensus quality score of sample i, $S_A(j)$ is the competence score of annotator j
(We assume that S_S and S_A are intrinsically aware of the annotator parameters a, w, b, and λ)
$\mathcal{E}$ defines the probability of exploring a new annotator

Output: New annotation(s) $\{k\}$ or $\{k, k'\}$

```
function RequestAnnotationExp($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, S_S(\cdot), S_A(\cdot), \mathcal{E}$)
    for all $i \in \mathcal{I}$ do
        $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$    ▷ Annotations of sample $i$
        $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$    ▷ Annotators of sample $i$
    end for
    for all $j \in \mathcal{J}$ do
        $\mathcal{K}^j \leftarrow \{k \in \mathcal{K} : j_k = j\}$    ▷ Annotations of annotator $j$
    end for
    $\mathcal{U}_s \leftarrow \{i \in \mathcal{I} : |\mathcal{K}_i| = 0\}$    ▷ Samples without any annotation
    $\mathcal{U}_a \leftarrow \{j \in \mathcal{J} : |\mathcal{K}^j| = 0\}$    ▷ Annotators without any annotation
    if $|\mathcal{U}_s| > 0$ then    ▷ If there is a sample without any annotation
        $i \leftarrow$ Randomly select from $\mathcal{U}_s$
    else
        $i \leftarrow \underset{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset}{\arg\min} S_S(i')$    ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample
    end if
    $\mathcal{R} \leftarrow \mathcal{U}_a \cap \mathcal{J}'$    ▷ Set of explorable annotators
    $\mathcal{T} \leftarrow \mathcal{J}' \setminus (\mathcal{J}_i \cup \mathcal{U}_a)$    ▷ Set of exploitable annotators
    if $|\mathcal{R}| > 0$ and $|\mathcal{T}| > 0$ then    ▷ If there are both explorable and exploitable annotators
        explore $\leftarrow$ true with probability $\mathcal{E}$    ▷ Randomly decide whether to explore a new annotator or exploit an existing annotator
    else if $|\mathcal{R}| > 0$ then    ▷ If there are only explorable annotators
        explore $\leftarrow$ true
    else if $|\mathcal{T}| > 0$ then    ▷ If there are only exploitable annotators
        explore $\leftarrow$ false
    end if
    if explore then
        $j \leftarrow$ Randomly select from $\mathcal{R}$    ▷ Select an annotator from explorable annotators
        $i' \leftarrow$ Randomly select from $\mathcal{I} \setminus \mathcal{U}_s$    ▷ Select a sample from previously annotated samples
        $k' \leftarrow$ Request an annotation for a random sample $i'$ from annotator $j$
    else
        $j \leftarrow \underset{j' \in \mathcal{J}' \setminus \mathcal{J}_i}{\arg\max} S_A(j')$    ▷ Select the most competent annotator from the set of active annotators who had not annotated sample $i$
    end if
    $k \leftarrow$ Request an annotation for the sample $i$ from annotator $j$
    if explore then
        return $\{k, k'\}$
    else
        return $\{k\}$
    end if
end function
```

O-CBS+ is based on Algorithm 1 with M-CBS as the EstimateLabels(·) function and RequestAnnotationExp(·) of Algorithm 3 as the RequestAnnotation(·) function. In this setting, RequestAnnotationExp(·) employs S_S as the sample consensus quality scoring function, same as O-CBS. Since in Section 4.3 we observe that O-CBS$\big(S_A^5\big)$ performs to our satisfaction, we fix the dominance suppression coefficient as $\varphi = 5$ and use S_A^5 as the annotator competence scoring function for O-CBS+. We denote O-CBS+ with different exploration parameters ($\mathcal{E}$) as O-CBS+($\mathcal{E}$). As a baseline method, we use O-CBS+(*Random*) which is similar to O-CBS(*Random*). In O-CBS+(*Random*), the annotators are selected randomly regardless of whether they are already known or new. Note that if there are samples without any annotation, the random selection is performed among them. As soon as all samples have annotations, full random selection commences.

In the remainder of this section, we thoroughly study the performance of O-CBS+. First, we investigate the effect of the exploration parameter $\mathcal{E}$ for all datasets and discuss the risks and benefits of incorporating new annotators into the system.. Then, we compare the performance of O-CBS+ with two methods (Welinder & Perona, 2010; Raykar & Agrawal, 2014) from the literature. Note that the work of Welinder and Perona provides the only directly comparable method to O-CBS+ as we have previously mentioned in Section 1.1. Raykar and Agrawal provide comparative results with the binary method of Welinder and Perona on the six Affective Text Analysis datasets using active crowd-labeling with binarized inputs. Although the method of Raykar and Agrawal is not directly comparable to our work, for the sake of completeness we also provide comparative results by binarizing our continuous-valued consensuses. Finally, we investigate the effect of enforcing a sample score related stopping criterion and provide further comparative results with Welinder and Perona, and Raykar and Agrawal.

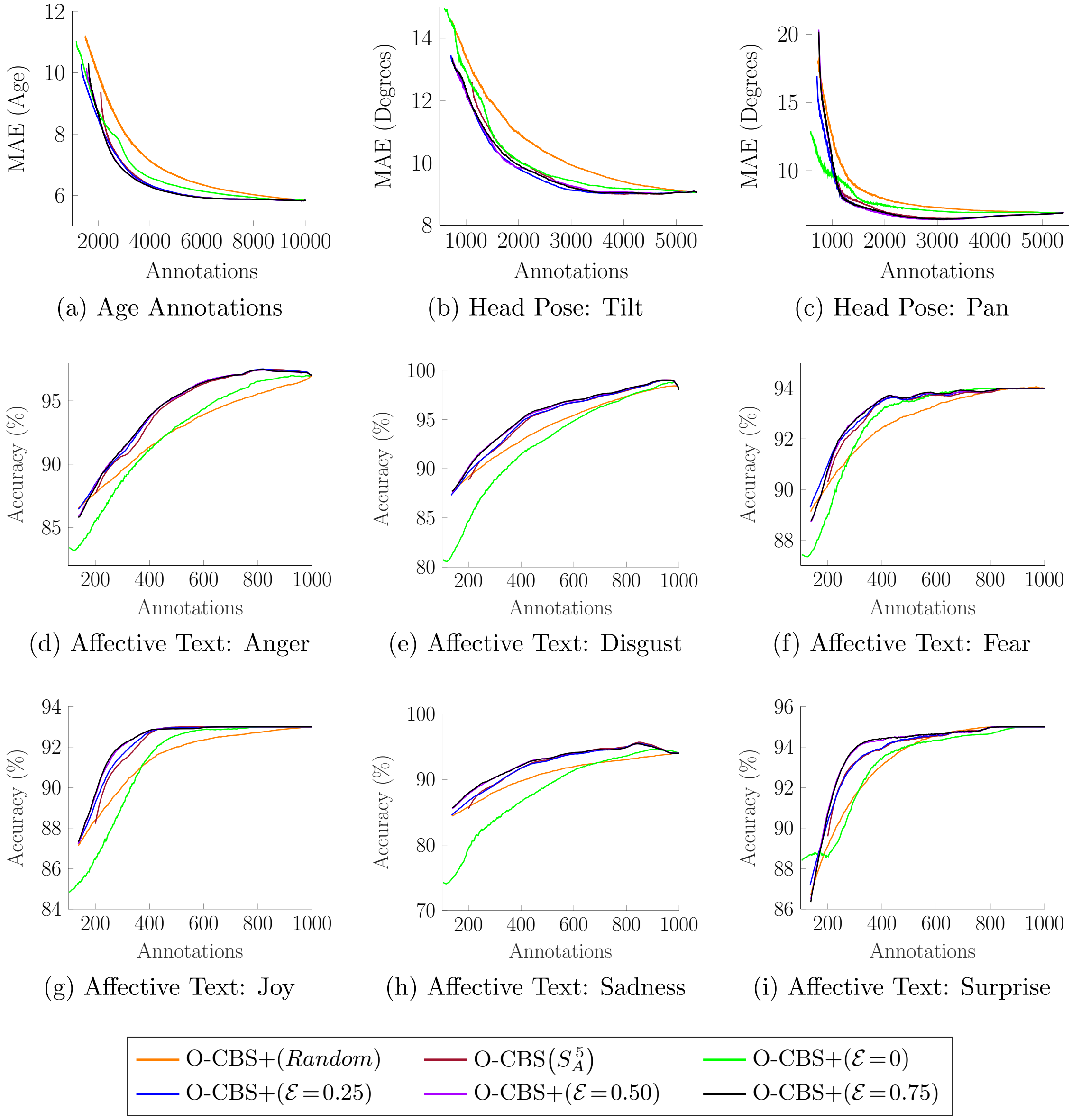
5.1 Effect of Annotator Exploration

In this section, we will discuss the results of starting active crowd-labeling from scratch under several different exploration parameters. We present our results on nine datasets, namely the Kara Age Annotations, the Head Pose Annotations, and the six Affective Text Analysis datasets.

5.1.1 Mean Absolute Age Error Improvement on the Kara Age Annotations Dataset

In Figure 8a, we present the effect of changing the exploration parameter $\mathcal{E}$ on the Kara Age Annotations dataset. Figure 8a shows the reduction in the mean absolute error in terms of age, while the active crowd-labeling is started from scratch. For the analysis to be meaningful, we start reporting the error once each sample has a consensus estimation. Therefore, the curves do not start from zero annotations. Additionally, due to the fact that the active crowd-labeling process has a random nature, the moment where every sample has a consensus is different for each trial. Thus, the starting point of the curves also differ from one another in the figures.

In Figure 8a, we compare O-CBS+ with fixed dominance suppression coefficient of $\varphi = 5$ for different $\mathcal{E}$ values. We also compare with O-CBS$\big(S_A^5\big)$ from Figure 6a and the random annotation selection mentioned in Section 4.3, as baseline comparisons. It is evident that



(a) Age Annotations

(b) Head Pose: Tilt

(c) Head Pose: Pan

(d) Affective Text: Anger

(e) Affective Text: Disgust

(f) Affective Text: Fear

(g) Affective Text: Joy

(h) Affective Text: Sadness

(i) Affective Text: Surprise

Figure 8: Effect of changing the exploration parameter $\mathcal{E}$ on the Kara Age Annotations, the Head Pose Annotations, and the Affective Text Analysis datasets. On the top row of figures we report the MAE and on the other rows we report the accuracy. The results are presented for $\varphi = 5$ and are the averages of 100 repetitions.

active learning from scratch with exploration performs better than the random selection method. We also observe that starting from scratch ensures the same success with fewer annotations.

An important point worth mentioning is that using O-CBS+($\mathcal{E}=0$) is not the same as using O-CBS with an empty set of initial annotations. Although $\mathcal{E} = 0$ seems like no exploration takes place in the process, inevitably exploration is done when there is no annotator to exploit. This case may also happen for any $\mathcal{E} < 1$. Similarly for $\mathcal{E} > 0$, when the system runs out of annotators to explore, it goes on full-exploitation mode until a new annotator joins the system.

When we observe Figure 8a, we see that the results get better and the gain eventually diminishes with higher exploration coefficient $\mathcal{E}$. Note that the annotator set is limited in the dataset, and thus the systems with large $\mathcal{E}$ values learn all annotators rapidly. When there are no new annotators to explore, the system begins to exploit high quality annotators early on. Therefore, better results are achieved faster. We have to keep in mind that the essence is in exploitation of high quality annotators, and this is achieved by exploration. Since the Kara Age Annotations dataset is a fairly large dataset, the difference between choosing different exploration coefficients quickly becomes indistinguishable after all annotators are explored. However, exploration should be used moderately on open ended annotation problems (i.e. where the annotator pool is considered to be unlimited).

5.1.2 Mean Absolute Degree Error Improvement on the Head Pose Annotations Datasets

Figures 8b and 8c show the effect of the exploration parameter $\mathcal{E}$ on the Head Pose Annotations *tilt* and *pan* datasets. Similar to the Kara Age Annotations dataset, we compare the O-CBS+ results with the two baseline methods O-CBS$\left(S_A^5\right)$ and O-CBS+(*Random*). On both datasets, increasing the exploration coefficient $\mathcal{E}$ results in marginal decrease in terms of mean absolute degree error. The results in Figures 8a to 8c suggest that the effect of $\mathcal{E}$ is difficult to observe on large datasets and call for a closer inspection on smaller datasets. The advantage of annotator selection over random selection is more apparent in the *tilt* dataset.

5.1.3 Accuracy Improvement on the Affective Text Analysis Datasets

In Figures 8d to 8i, we present the effect of the exploration parameter $\mathcal{E}$ on the Affective Text Analysis datasets, which are significantly smaller datasets compared to the other three datasets. Overall, the results are in concord with those of the Kara Age Annotations dataset (Figure 8a) and the Head Pose Annotations dataset (Figures 8b and 8c). In addition, the advantage of using a higher exploration parameter such as $\mathcal{E} = 0.75$ results in higher accuracies.

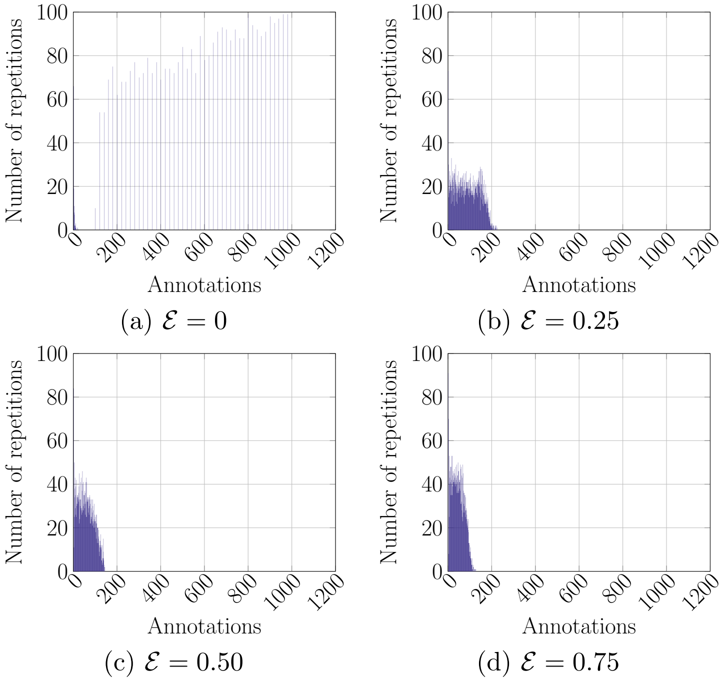
Since the annotation set is limited, all curves converge to the same point toward the end of the active crowd-labeling process. Therefore, well-performing methods which reach a higher accuracy with fewer annotations converge to the same point with the weaker methods at the end. An example for this can be observed in Figure 8i, where the exploration-based methods outperform O-CBS$\left(S_A^5\right)$ but end up with the same accuracy at the end.

A striking difference from the Kara Age Annotations dataset is the performance of the $\mathcal{E} = 0$ curve. In the six Affective Text Analysis datasets, it significantly falls behind its counterparts. The strict imposition of annotator exploitation results in the late integration of high-quality annotators to the system. Since the Affective Text Analysis datasets are much smaller than the Kara Age Annotations dataset, timely exploration of high-quality annotators is much more critical for the success of the active learning process and the tardiness caused by selecting $\mathcal{E} = 0$ becomes evident in the graphs.

On specifically three datasets, namely *fear*, *joy*, and *surprise*, our method quickly reaches high accuracies with a small number of annotations. This is due to the fact that our method succeeds in selecting high-quality annotators faster. Another remark is about the peaks observed in the *anger*, *disgust*, and *sadness* datasets. These peaks indicate that the system has to exploit low-quality annotators when it runs out of annotations from the high-quality ones. The reason is that we are working with a limited annotation set and we force the system to use every annotation for observing the complete behavior. Therefore, the active learning performance degrades in these three datasets with an increasing number of annotations toward the end.

5.2 Is It Wise to Take Risks by Incorporating New Annotators?

Although it is apparent that a system without exploration would suffer when the starting annotation set is small, the intuitive expectation is that a conservative approach to exploration would be better. This is due to the fact that there is a risk associated with new annotators and we can always select the better annotators among the annotators we know. However, the results in Section 5.1 show otherwise.



(a) $\mathcal{E} = 0$ (b) $\mathcal{E} = 0.25$

(c) $\mathcal{E} = 0.50$ (d) $\mathcal{E} = 0.75$

Figure 9: New annotator exploration times on the Affective Text Analysis - Anger dataset for O-CBS+$\left(S_A^5\right)$

When we observe the exploration times shown in Figure 9, we see that the system exhausts new annotators quickly since our datasets contain finite number of annotators. When working with a limited annotator set, it is wise to assess all annotators quickly so that the active crowd-labeling approach starts to utilize better annotators early on. The results presented in Figure 9 and Section 5.1 validate this observation. A larger $\mathcal{E}$ results in the addition and assessment of new annotators to the system very quickly and therefore better results are achieved with fewer annotations by utilizing good annotators.

Note that these results are obtained from readily available datasets with a limited number of annotators. In a live and open-ended active crowd-labeling process, it would be wise to concentrate more on exploiting the existing good annotators and choose a smaller $\mathcal{E}$ value, instead of constantly exploring new annotators.

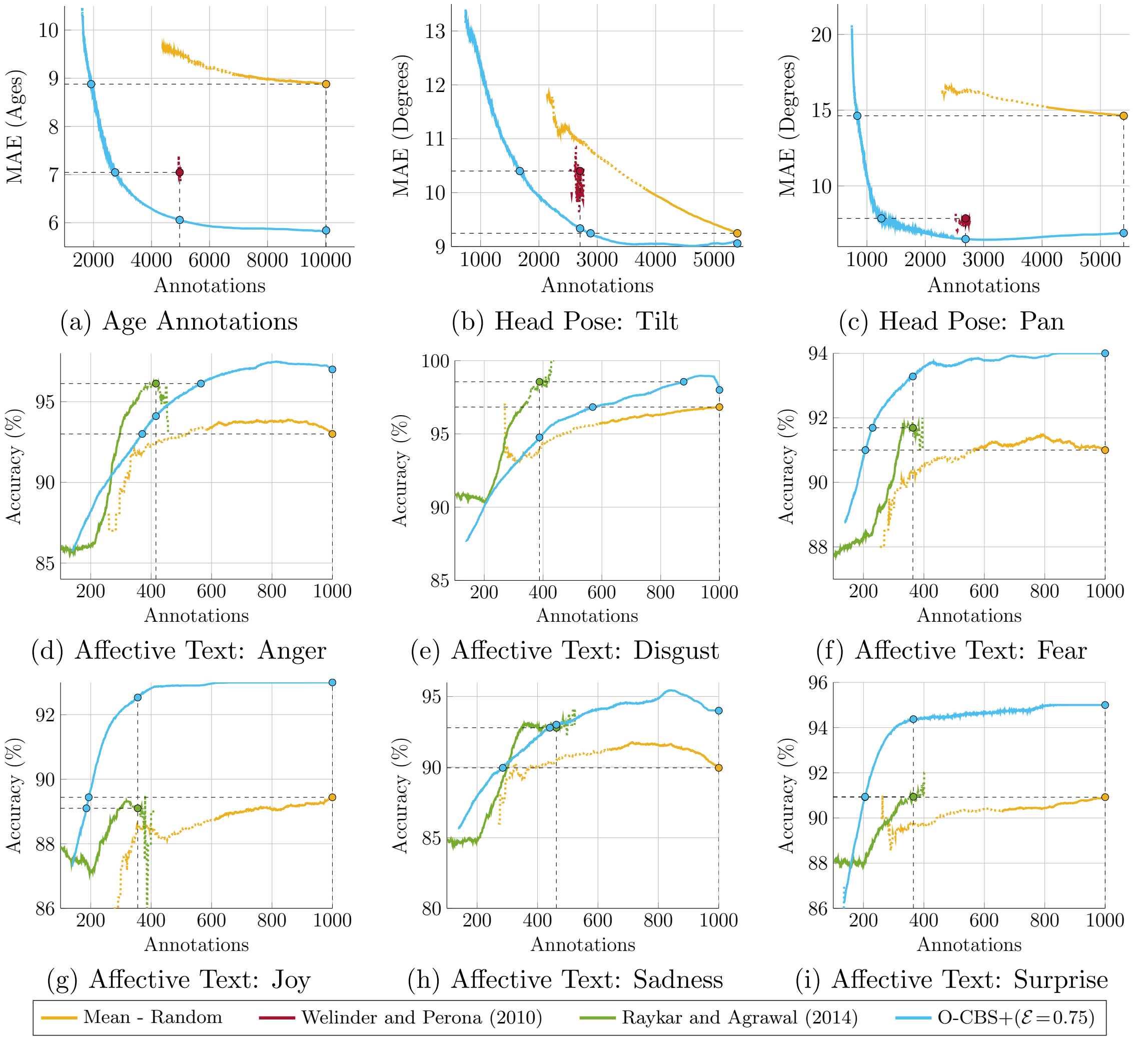
5.3 Comparative Performance of O-CBS+ Under Annotation Count Limitations

So far, we have deduced that S_A^5 is a good annotator competence scoring function choice and fixed it in O-CBS+. Figure 8 shows that fast exploration of annotators is preferable, especially for small datasets. Thus, we present the results using O-CBS+($\mathcal{E}=0.75$) for the comparative performance evaluation of O-CBS+ with the existing methods, namely Welinder and Perona (2010), and Raykar and Agrawal (2014). The experiments with both opponent methods and our method O-CBS+($\mathcal{E}=0.75$) are repeated 100 times.

In Figure 10, we compare our method with the Mean-Random baseline method and the method of Welinder and Perona (2010) on the Kara Age Annotations and Head Pose Annotations datasets, and the method of Raykar and Agrawal (2014) on the Affective Text Analysis datasets. By the very nature of active crowd-labeling, annotations of the samples are acquired gradually. Thus, in the early steps of the process, not every sample has an estimated label. Moreover, the required number of annotations for obtaining consensus label of every sample varies depending on the sample selection strategy of the method in question. However, for the mean absolute error (MAE) and accuracy comparisons to make sense, every sample's consensus error must contribute to the mean. For this reason, we represent the initial part of the process where some sample labels do not have estimations by dotted lines in the plots. Additionally, both methods by Welinder and Perona, and Raykar and Agrawal employ stopping criteria which results in the algorithms stopping at different annotation counts among 100 repetitions. Therefore, the ends of the curves are also shown in dotted lines when the MAE or the accuracy is calculated with fewer than 100 repetitions. The middle portions of the curves are shown in solid lines.

In the Mean-Random baseline method, the annotations are added randomly and the mean of the annotations of a sample are used as the resulting label. In Figure 10a, we observe that the mean absolute age error achieved by this baseline method on the Kara Age Annotations dataset using all 10020 annotations can be matched by O-CBS+ with 1796 annotations (∼18% of all annotations). Figures 10b and 10c show that our method can match the performance of the Mean-Random baseline method on the Head Pose Annotations tilt and pan datasets with 2886 and 836 annotations (∼53% and ∼15% of all annotations), respectively.

Figures 10d to 10i present the performance of O-CBS+($\mathcal{E}=0.75$) against the method of Raykar and Agrawal (2014) on the Affective Text Analysis datasets, accompanied with the Mean-Random method as the baseline. Similar to Figures 10a to 10c, O-CBS+($\mathcal{E}=0.75$) outperforms the Mean-Random method across all six datasets. Our method matches the end result of the Mean-Random method, using a minimum of 193 and a maximum of 569



(a) Age Annotations

(b) Head Pose: Tilt

(c) Head Pose: Pan

(d) Affective Text: Anger

(e) Affective Text: Disgust

(f) Affective Text: Fear

(g) Affective Text: Joy

(h) Affective Text: Sadness

(i) Affective Text: Surprise

Figure 10: Comparison of O-CBS+ with the method of Welinder and Perona (2010) on the Kara Age Annotations and the Head Pose Annotations datasets and with the method of Raykar and Agrawal (2014) on the Affective Text Analysis datasets. On the top row of figures we report the MAE and on the other rows we report the accuracy. The plots are presented when at least one out of 100 repetitions has annotations for all samples. The dotted lines continue as solid lines whenever all 100 repetitions have annotations for all samples. The circles mark the required annotation counts for our method to reach the performances of baseline Mean-Random method and the contender methods. The horizontal black dashed lines provide visual guide.

annotations across the six datasets, and thereby resulting in a ∼70% cost reduction on average.

We support the findings of Figure 10 with a more detailed breakdown of the comparative results, presented in Table 4. We perform t-test for validating the statistical significance of the results presented in Figure 10. For comparison, we take the number of

	Welinder and Perona (2010)		O-CBS+($\mathcal{E}=0.75$)	Required annotations for O-CBS+($\mathcal{E}=0.75$) to reach target MAE
Dataset	Annotations	MAE	MAE at target annotations	
Age	4969.77	7.02 ages	**6.06** ages	**2775.98**
Tilt	2705.03	10.10 degrees	**9.33** degrees	**1892.16**
Pan	2689.77	7.58 degrees	**6.49** degrees	**1387.88**

(a) Comparison with Welinder and Perona (2010) on the Kara Age Annotations and the Head Pose Annotations datasets

	Raykar and Agrawal (2014)		O-CBS+($\mathcal{E}=0.75$)	Required annotations for O-CBS+($\mathcal{E}=0.75$) to reach target accuracy
Dataset	Annotations	Accuracy (%)	Accuracy at target annotations (%)	
Anger	415.86	96.07	*94.11*	*535.81*
Disgust	387.78	98.92	*94.76*	*726.82*
Fear	363.49	91.50	**93.28**	**247.32**
Joy	355.51	89.17	**92.53**	**196.22**
Sadness	462.34	93.31	**93.01**	*522.80*
Surprise	365.22	91.60	**94.38**	**231.41**

(b) Comparison with Raykar and Agrawal (2014) on the Affective Text Analysis datasets

Table 4: The effect of enforcing annotation count or MAE/accuracy limit and the comparative performance results on the Kara Age Annotations, the Head Pose Annotations, and the Affective Text Analysis datasets. The tables indicate the results of the t-test with significance level 0.01 across 100 repetitions, using bold underlined font when our method performs **better**, bold font when the test is **inconclusive**, and italic font when our method performs *worse*.

annotations at which an opponent algorithm stops, and use this as a stopping criterion for O-CBS+($\mathcal{E}=0.75$) to report the MAE or accuracy. Additionally, we also take the MAE or accuracy at which an opponent algorithm stops, and report the mean number of annotations needed to reach this target using O-CBS+($\mathcal{E}=0.75$). Significance test results against opponent methods are reported under the rightmost two columns, where the underlined bold values indicate that our method is significantly superior than the opponent method. Bold values indicate a tie and italic values indicate that the opponent method is better. The results for the opponent methods are given in regular script as reference values.

In Table 4a, we observe the significance test results of O-CBS+($\mathcal{E}=0.75$) against the method of Welinder and Perona (2010). On the Kara Age Annotations dataset, the algorithm of Welinder and Perona stops at 4970 annotations on average and a little more than half on the annotations are unused because they come from annotators marked as spammers. At this point, the lowest mean absolute error is reached. For matching the same MAE, our method requires 2776 annotations on average, and achieves better overall performance as more annotators are employed. Similar results are also observed for the the Head

Pose Annotations tilt and pan datasets, where O-CBS+ proves to be an effective algorithm both in terms of achieving significantly lower error with annotation count limitations and by using significantly fewer annotations for a targeted MAE.

Note that Welinder and Perona (2010) do not employ sample prioritization. They acquire annotations for each sample one by one. For each sample, they acquire as many annotations as they can and move onto the next sample. Thus, the point where each sample has a consensus value occurs later in the annotation process. This is why in Figures 10a to 10c the red curves preceding the red dots are almost invisible since the annotation acquisition process stops after a very short while. Once their algorithm flags an annotator as a spammer, that annotator is not consulted anymore.

Compared to the method of Welinder and Perona (2010), our method uses a more complex scheme. First, we employ sample prioritization by sample consensus quality scoring. Second, instead of grouping the annotators into two discrete groups as spammers and non-spammers, we rank them according to four parameters for each annotator. This way, better annotators are also ranked among themselves while low-quality annotators are ignored until the end of the annotation process. Low-quality annotators may also be completely excluded from the annotation process by a simple thresholding mechanism on the annotator competence score.

An additional observation about these methods' performances on the *tilt* dataset is that the algorithm of Welinder and Perona (2010) falls short of achieving the Mean-Random baseline method's performance. This is due to the fact that many annotators are marked as spammers and the annotation process stops very early. Another reason is that the *tilt* dataset is actually quite a challenging dataset in the sense that the baseline method achieves a close performance to our method O-CBS+($\mathcal{E}=0.75$), albeit using all annotations.

In Figures 10d to 10i and Table 4b, we present the performance of O-CBS+($\mathcal{E}=0.75$) against the method of Raykar and Agrawal (2014) on the Affective Text Analysis datasets. In contrast to Welinder and Perona (2010), Raykar and Agrawal employ a more intricate annotation selection algorithm and the change in the accuracy over time (the green lines in Figures 10d to 10i) is observable since all samples have annotations. Our method succeeds to achieve higher accuracies at the targeted number of annotations in the *fear*, *joy*, and *surprise* datasets with a significant margin and is tied on the *sadness* dataset. Although our method seems to struggle in the *anger* and *disgust* datasets, observing Figures 10d and 10e shows that the overall performance of our method in the long run (i.e. without annotation count limit) is capable of achieving a higher or similar accuracy. These findings confirm that O-CBS+ is overall a better approach to the active crowd-labeling problem with significant gains on annotation expenses.

5.4 Comparative Performance of O-CBS+ While Enforcing a Sample Score Related Stopping Criterion

In Equation 6, S_S is defined as the precision (reciprocal of the variance) of the posterior distribution of the sample consensus. In both O-CBS and O-CBS+, the aim is to reduce this variance value (i.e. increase S_S) for each sample. The algorithms are designed to choose the sample with the lowest S_S to be annotated in each annotation step. Thus, the overall direction is the enhancement of every sample's score (i.e. reducing the sample consensus posterior variance) during the course of active crowd-labeling.

So far, we were not concerned with the question of how high S_S should be for having a satisfactory sample consensus. Our aim was to increase consensus quality as much as possible within the annotation budget limit. In Figures 4, 6, 8 and 10, we show the performance of the proposed methods with only the budget limit as an enforceable stopping criterion. Every point on those graphs actually show the performance of the corresponding method for every possible annotation budget limit. However, this approach does not consider the adequacy of sample consensus values, and is at risk of prematurely ending the active crowd-labeling process or overspending by collecting excessive annotations.

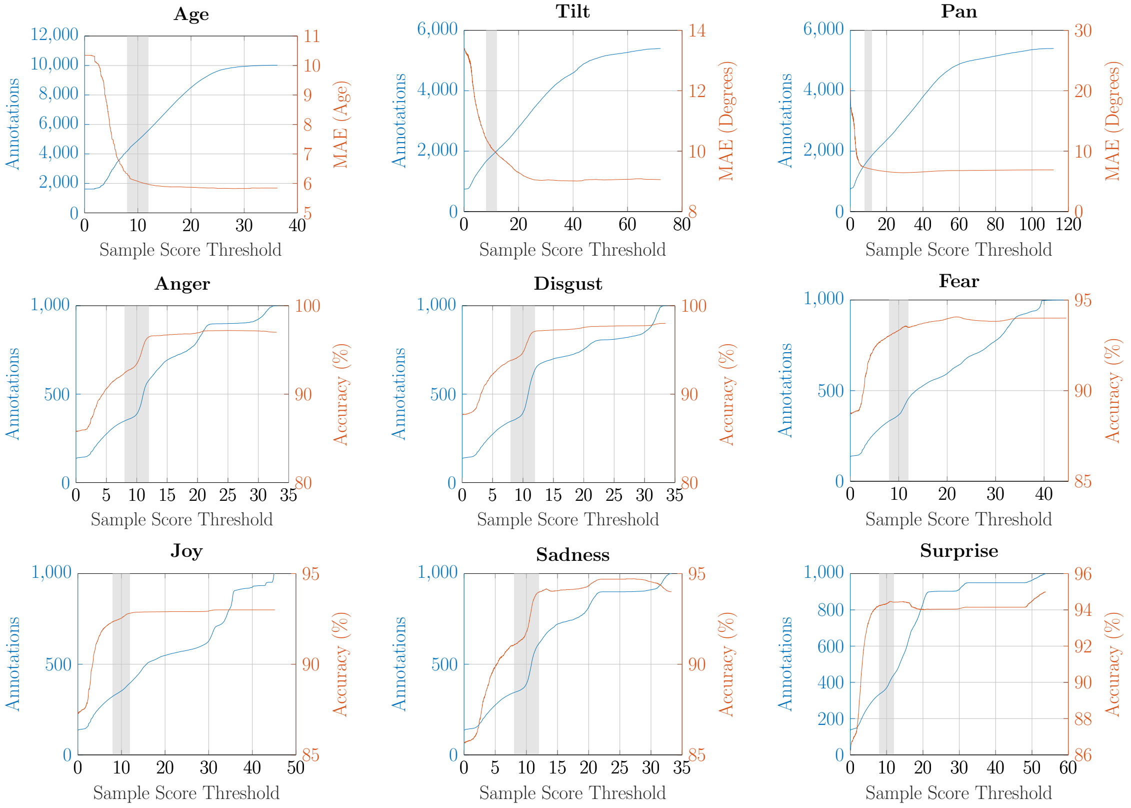


Figure 11: The effect of enforcing the sample scoring threshold τ on the number of annotations and model performance. Blue curves show the final annotation count (i.e. cost) when τ is enforced and red curves show the performance at the end of the annotation process for a given τ. On the top row of figures we report the MAE and on the other rows we report the accuracy as indicators of performance. The gray bands in the plots show the region between $\tau = 8$ and $\tau = 12$. The results are reported for the average of 100 repetitions.

To address this concern, we aim to stop the annotation process upon attaining satisfactory sample consensus values for all samples by setting a target on the sample consensus posterior variance, namely δ. This is equivalent to stopping the active crowd-labeling process when every sample has a satisfactory score S_S, i.e.

$$\min_i S_S(i) > \underbrace{\frac{1}{\delta}}_{\tau}, \tag{11}$$

since S_S is the reciprocal of the posterior variance. Therefore, τ signifies the target lower limit on S_S.

The cost associated with the active crowd-labeling systems consists of not only the annotation budget, but also the cost of reaching erroneous consensuses (which may also have monetary repercussions). System designers are often faced with making a trade-off between performance and budget to find a sensible operation range. In our case, collecting more annotations often result in reduced error while increasing expenses. Due to the nature of the sample score S_S and O-CBS+, choosing a high τ value would result in lower error and is preferable if the cost of making error is high. In contrast, system designers working with very limited budgets may resort to using a lower τ value. A reasonably low value for the posterior variance of a sample's consensus is 0.1. Enforcing a stopping criterion to reach this goal for each sample corresponds to choosing $\tau = 10$.

In Figure 11, we show the performance of O-CBS+($\mathcal{E}=0.75$) for varying τ values. Blue curves show the final annotation count (i.e. cost) when τ is enforced and red curves show the performance at the end of the annotation process for a given τ. The gray bands in the plots show the region around $\tau = 10$; specifically, the bands rest between $\tau = 8$ and $\tau = 12$. The plots show promising performance and annotation count values inside the gray bands. The results verify our previous deductions. Especially, for *anger*, *disgust*, and *sadness* datasets where our methods struggle, $\tau = 10$ presents a turning point for both error and budget. Additionally, in the remaining datasets the gray band areas signify very preferable operation ranges.

In Table 5, we give the results of O-CBS+($\mathcal{E}=0.75$) for different τ values compared to the methods of Welinder and Perona (2010), and Raykar and Agrawal (2014). The experiments with both the opponent methods and our method O-CBS+($\mathcal{E}=0.75$) are repeated 100 times. We perform t-test for validating the statistical significance of the results. We report the number of annotations and the error/accuracy when our algorithm stops for the τ values 8, 10, and 12. Significance test results against opponent methods are reported under the O-CBS+($\mathcal{E}=0.75$) heading, where underlined bold values indicate that our method is significantly superior than the opponent method. Bold values indicate a tie and italic values indicate that the opponent method is better. The results for the opponent methods are given in regular script as reference values.

The results show that for $\tau = 8$, the number or annotations at which our algorithm stops are always significantly lower than its contenders, with acceptable error or accuracy values. When $\tau = 10$, our algorithm is tied with or better than its contenders in terms of annotation count and the accuracies improve, especially for the *tilt*, *anger*, *disgust*, and *sadness* datasets. For $\tau = 12$, our algorithm achieves significantly superior performance across all datasets except *disgust* in terms of error/accuracy at the expense of increasing cost.

	Welinder and Perona (2010)		O-CBS+($\mathcal{E}=0.75$)					
			$\tau = 8$		$\tau = 10$		$\tau = 12$	
Dataset	Annotations	MAE	Annotations	MAE	Annotations	MAE	Annotations	MAE
Age	4969.77	7.02	**4189.93**	**6.33**	**4911.37**	**6.07**	*5607.13*	**5.97**
Tilt	2705.03	10.10	**1657.70**	*10.42*	**1836.39**	**10.11**	**2009.94**	**9.92**
Pan	2689.77	7.58	**1560.16**	**7.32**	**1721.22**	**7.13**	**1868.02**	**7.01**

(a) Comparison with Welinder and Perona (2010) on the Kara Age Annotations and the Head Pose Annotations datasets

	Raykar and Agrawal (2014)		O-CBS+($\mathcal{E}=0.75$)					
			$\tau = 8$		$\tau = 10$		$\tau = 12$	
Dataset	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)
Anger	415.86	96.07	**347.83**	*93.38*	**386.20**	*94.58*	*564.59*	**97.24**
Disgust	387.78	98.92	**346.12**	*94.64*	**392.72**	*95.53*	*625.24*	*97.41*
Fear	363.49	91.50	**331.49**	**93.45**	**365.74**	**93.77**	*458.29*	**93.74**
Joy	355.51	89.17	**323.10**	**92.59**	**352.96**	**92.79**	*394.22*	**92.98**
Sadness	462.34	93.31	**343.58**	*91.96*	**390.84**	*92.72*	*603.89*	**94.50**
Surprise	365.22	91.60	**334.87**	**94.60**	**371.00**	**94.67**	*447.00*	**94.64**

(b) Comparison with Raykar and Agrawal (2014) on the Affective Text Analysis datasets

Table 5: The effect of enforcing different stopping criteria and the comparative performance results on the Kara Age Annotations, the Head Pose Annotations, and the Affective Text Analysis datasets. The tables indicate the results of the t-test with significance level 0.01 across 100 repetitions, using bold underlined font when our method performs **better**, bold font when the test is **inconclusive**, and italic font when our method performs *worse*.

6. Conclusions and Future Directions

In this work, we introduce two active crowd-labeling algorithms for the crowdsourced labeling process, namely O-CBS and O-CBS+. We base our methods on selecting the most beneficial annotation by determining annotator and sample consensus qualities. In addition to a novel sample consensus quality score, we also introduce a family of competence scoring functions designed to prevent annotator domination. Both O-CBS and O-CBS+ are capable of utilizing a wide range of sample consensus quality and annotator competence scoring functions, inclusive of the two novel approaches we introduce.

We investigate the effect of the dominance suppression factor and annotator exploration/exploitation trade-off over nine different real-world datasets. A thorough investigation of the dominance suppression factor in the annotator competence scoring function reveals that preventing annotator domination is of utmost importance in assessing the annotator quality correctly. The results also indicate that the timely exploration of new annotators is crucial for high quality consensus estimation. Additionally, we reduce the computational cost of the consensus estimation phase in the active crowd-labeling process, which constitutes a significant portion of the total CPU time.

We introduce the Head Pose Annotations datasets with *tilt* and *pan* attributes and test O-CBS+ on these datasets, in addition to the publicly available Kara Age Annotations and Affective Text Analysis datasets. Our method measures up to and surpasses the literature standards by using as few as one fifth of the annotations (i.e. $\sim 80\%$ cost reduction). We also investigate a sample score related stopping criterion so that the active crowd-labeling process is terminated automatically when the sample consensuses exceed an acceptable quality.

In some annotation problems, annotators are asked to annotate multiple attributes for a single sample. This is the case for the Head pose Annotations and the Affective Text Analysis datasets, which have two and six attributes, respectively. In this work, we handle the annotations of each attribute as separate and independent datasets. However, it could be beneficial to use those attributes together for understanding the behavior of the annotator better. Investigating the dependencies between the annotations of different attributes and introducing a multivariate model of sample consensus are left to be explored as promising future works. Relaxing the homogeneous sample difficulty assumption by incorporating a heterogeneous sample difficulty parameter is another interesting future direction. Additionally, it may be worthwhile to investigate the effects of different sample consensus quality and annotator competence scoring functions on the active crowd-labeling system. Furthermore, addressing the issue of annotator competence fluctuation over time and distributing the tasks according to the recent performance of the annotators is also left to be explored in the future.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editors for their invaluable comments and constructive criticisms. Their efforts helped to improve the technical quality and presentation of this manuscript immensely. The authors would also like to thank Dr. Çetin Meriçli for his help in proofreading the manuscript. This work is supported by the Ministry of Development of the Republic of Turkey under the project TAM with the project number 2007K120610.

Appendix A. Deriving the Posterior Distribution of the Sample Consensus

Proposition 1. *If the distribution of* y_k *is* $\mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right)$, *then the posterior distribution of* x_i *is*

$$x_i | \{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} (w_{j_k}^{-1} a_{j_k} y_k - b_{j_k})}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum_{k:i_k=i} \lambda_{j_k} \right)^{-1} \right) \tag{12}$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ *is the set of parameters of annotator* j *and* $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ *is the set of annotations of sample* i.

Proof. Let N, R, and K be number of samples, annotators, and annotations, respectively.

$$p(y_{1:K}|x_{1:N},\theta_{1:R}) = \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \tag{13}$$

$$\begin{aligned}
\log p(y_{1:K}|x_{1:N},\theta_{1:R}) &= \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) && (14)\\
&= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) && (15)\\
&= \sum_{k=1}^{K}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\left(\frac{(y_k - a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}))^2}{\frac{w_{j_k}^2}{\lambda_{j_k}}}\right)\right) && (16)\\
&= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\left(\lambda_{j_k}\frac{(y_k - a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}))^2}{w_{j_k}^2}\right) && (17)\\
&= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1}y_k - a_{j_k}(x_{i_k}+b_{j_k}))^2 && (18)
\end{aligned}$$

Since $a_j \in \{-1,1\}\ \forall j$, $a_j^2 = 1$:

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k} - x_{i_k})^2 \tag{19}$$

From Bayes' rule we know that

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) = \frac{p(y_{1:K}|x_{1:N},\theta_{1:R})p(x_i)}{p(y_{1:K}|\theta_{1:R})} \tag{20}$$

Since the prior of x_i is flat

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) \propto p(y_{1:K}|x_{1:N},\theta_{1:R}) \tag{21}$$

By omitting independent variables, we get

$$p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) \propto p(y_{1:K}|x_{1:N},\theta_{1:R}) \tag{22}$$

Combining Equations 19 and 22 gives us

$$\log p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) \propto -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k} - x_{i_k})^2 \tag{23}$$

By omitting the terms without x_i we get:

$$\begin{aligned}
&\propto -\frac{1}{2}\sum_{k:i_k=i} \lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k} - x_i)^2 && (24)\\
&\propto -\frac{1}{2}\sum_{k:i_k=i} \Big(\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})^2 + \lambda_{j_k}x_i^2 \\
&\qquad\qquad -2x_i\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})\Big) && (25)
\end{aligned}$$

Rearranging and omitting the terms without x_i:

$$\propto -\frac{1}{2}x_i^2 \underbrace{\sum_{k:i_k=i}\lambda_{j_k}}_{\sigma^{-2}} + x_i \underbrace{\sum_{k:i_k=i}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})}_{\mu\sigma^{-2}} \tag{26}$$

The equation is in the form of Normal distribution. Therefore, we have

$$p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) = \mathcal{N}\left(x_i; \frac{\sum_{k:i_k=i}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})}{\sum_{k:i_k=i}\lambda_{j_k}}, \left(\sum_{k:i_k=i}\lambda_{j_k}\right)^{-1}\right) \tag{27}$$

□

Appendix B. Starting Subset Creation Algorithm for Active Crowd-Labeling

Algorithm 4 gives the details of the starting subset creation process. We aim to create starting subsets from the annotation data that satisfy the following conditions:

- The resulting subset should have ν annotations,
- Minimum sample count of the resulting subset should be ρ,
- Minimum annotator count of the resulting subset should be η,
- Every annotator in the resulting subset should have at least ζ annotations,
- Every sample in the resulting subset should have at least δ annotations,
- Annotations of an annotator should not be disconnected from the rest of the data. Every annotator must have an annotation for a sample that also has an annotation from another annotator.

Algorithm 4 Create Starting Set By Elimination

Input:

Sets of samples $\mathcal{I}$, annotators $\mathcal{J}$, annotations $\mathcal{K}$

Target annotation count ν, minimum annotations per annotator ζ, minimum annotations per sample δ, minimum sample count ρ, minimum annotator count η

Output:

Subset of annotations $\mathcal{K}$

```
function CreateSubset($\mathcal{I}, \mathcal{J}, \mathcal{K}, \nu, \zeta, \delta, \rho, \eta$)
    Shuffle($\mathcal{K}$)
    for all $k \in \mathcal{K}$ do
        for all $j \in \mathcal{J}$ do
            $\mathcal{K}^j \leftarrow \{k \in \mathcal{K} : j_k = j\}$    ▷ Annotations of the annotator $j$
        end for
        for all $i \in \mathcal{I}$ do
            $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$    ▷ Annotations of the sample $i$
        end for
        if $|\mathcal{K}^{j_k}| < \zeta$ then    ▷ If the annotator $j_k$ of the annotation $k$ has less than $\zeta$ annotations
            $\mathcal{D} \leftarrow \mathcal{K}^{j_k}$    ▷ Mark all annotations of $j_k$ to be removed
        else
            $\mathcal{D} \leftarrow \{k\}$    ▷ Mark only the annotation $k$ to be removed
        end if
        $\mathcal{T}_s \leftarrow \{i \in \mathcal{I} : |\mathcal{K}_i \setminus \mathcal{D}| > 0\}$    ▷ Samples with at least 1 annotation
        $\mathcal{T}_a \leftarrow \{j \in \mathcal{J} : |\mathcal{K}^j \setminus \mathcal{D}| > 0\}$    ▷ Annotators with at least 1 annotation
        if $\exists i \in \{i_k : k \in \mathcal{D}\}$ s.t. $|\mathcal{K}_i \setminus \mathcal{D}| < \delta$ then    ▷ If any sample has less than $\delta$ annotations
            continue    ▷ Reject
        else if $|\mathcal{T}_s| < \eta$ or $|\mathcal{T}_a| < \rho$ then    ▷ If number of samples or annotators are below limits
            continue    ▷ Reject
        else if $\exists j$ s.t. $|\mathcal{K}_i \setminus \mathcal{D}| = 1, \forall i \in \mathcal{I}_j$ then    ▷ If an annotator does not have a common sample annotated with another annotator
            continue    ▷ Reject
        else    ▷ Accept the removal of the annotation(s) in $\mathcal{D}$
            $\mathcal{K} \leftarrow \mathcal{K} \setminus \mathcal{D}$    ▷ Update $\mathcal{K}$ by removing $\mathcal{D}$
        end if
        if $|\mathcal{K}| < \nu$ then    ▷ Break if target annotation count is reached
            break
        end if
    end for
    return $\mathcal{K}$
end function
```

References

Dawid, A. P., & Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 20–28.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Donmez, P., & Carbonell, J. G. (2008a). Paired-Sampling in Density-Sensitive Active Learning. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*.

Donmez, P., & Carbonell, J. G. (2008b). Proactive Learning: Cost-Sensitive Active Learning with Multiple Imperfect Oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 619–628.

Donmez, P., Carbonell, J. G., & Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09*, p. 259. ACM Press.

Fang, M., Yin, J., & Tao, D. (2014). Active learning for crowdsourcing using knowledge transfer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pp. 1809–1815. AAAI Press.

Frnay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869.

Fu, Y., Zhu, X., & Li, B. (2013). A survey on instance selection for active learning. *Knowledge and Information Systems*, *35*(2), 249–283.

Gao, J., Liu, X., Ooi, B. C., Wang, H., & Chen, G. (2013). An online cost sensitive decision-making method in crowdsourcing systems. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pp. 217–228. ACM.

Gourier, N., Hall, D., & Crowley, J. L. (2004). Estimating face orientation from robust detection of salient facial structures. In *Proceedings of the FGNet Workshop on Visual Observation of Deictic Gestures*, pp. 1–9. FGnet (IST–2000–26434) Cambridge, UK.

Guo, S., Parameswaran, A., & Garcia-Molina, H. (2012). So who won?: Dynamic max discovery with the crowd. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pp. 385–396. ACM.

Ho, C.-J., Jabbari, S., & Vaughan, J. W. (2013). Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. I–534–I–542. JMLR.org.

Ho, C.-J., Slivkins, A., & Vaughan, J. W. (2016). Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, *55*, 317–359.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009*

Workshop on Active Learning for Natural Language Processing, HLT '09, pp. 27–35. Association for Computational Linguistics.

Jagabathula, S., Subramanian, L., & Venkataraman, A. (2014). Reputation-based worker filtering in crowdsourcing. In *Advances in Neural Information Processing Systems 27*, pp. 2492–2500. Curran Associates, Inc.

Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, pp. 467–474.

Kamar, E., Kapoor, A., & Horvitz, E. (2013). Lifelong Learning for Acquiring the Wisdom of the Crowd. *International Joint Conference on Artificial Intelligence (IJCAI)*, *13*, 2313–2320.

Kamar, E., Kapoor, A., & Horvitz, E. (2015). Identifying and Accounting for Task-Dependent Bias in Crowdsourcing. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)*, pp. 92–101.

Kara, Y. E., Genc, G., Aran, O., & Akarun, L. (2015). Modeling annotator behaviors for crowd labeling. *Neurocomputing*, *160*, 141–156.

Karger, D. R., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems 24*, pp. 1953–1961. Curran Associates, Inc.

Karger, D. R., Oh, S., & Shah, D. (2014). Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *Operations Research*, *62*(1), 1–24.

Khetan, A., & Oh, S. (2016). Achieving budget-optimality with adaptive schemes in crowdsourcing. In *Advances in Neural Information Processing Systems 29*, pp. 4844–4852. Curran Associates, Inc.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., & Murphy, K. (2017). Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*.

Lakshminarayanan, B., & Teh, Y. (2013). Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *arXiv preprint arXiv:1305.0015*, 1–19.

Li, H., Zhao, B., & Fuxman, A. (2014). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pp. 165–176.

Li, Q., Ma, F., Gao, J., Su, L., & Quinn, C. J. (2016). Crowdsourcing High Quality Labels with a Tight Budget. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*, pp. 237–246.

Lin, C. H., Mausam, & Weld, D. S. (2016). Re-active learning: Active learning with relabeling. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 1845–1852. AAAI Press.

Marcus, A., Karger, D., Madden, S., Miller, R., & Oh, S. (2013). Counting with the crowd. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pp. 109–120. VLDB Endowment.

Mozafari, B., Sarkar, P., Franklin, M., Jordan, M., & Madden, S. (2014). Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning. In *Proceedings of the VLDB Endowment*, Vol. 8, pp. 125–136.

Nguyen, A. T., Wallace, B. C., & Lease, M. (2015). Combining Crowd and Expert Labels using Decision Theoretic Active Learning. In *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*, pp. 120–129.

Ok, J., Oh, S., Shin, J., Jang, Y., & Yi, Y. (2017). Iterative Bayesian Learning for Crowd-sourced Regression. *arXiv:1702.08840 [cs.LG]*, 1–22.

Peng, J., Liu, Q., Ihler, A., & Berger, B. (2013). Crowdsourcing for structured labeling with applications to protein folding. In *Proceedings of the ICML Workshop on Machine Learning Meets Crowdsourcing*, pp. 2008–2012.

Raykar, V. C., & Agrawal, P. (2014). Sequential crowdsourced labeling as an epsilon-greedy exploration in a Markov Decision Process. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, Vol. 33, pp. 832–840.

Raykar, V. C., & Yu, S. (2011). Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing Systems 24*, pp. 1809–1817. Curran Associates, Inc.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *99*, 1297–1322.

Settles, B. (2010). Active Learning Literature Survey. Tech. rep. 1648, University of Wisconsin-Madison, Computer Sciences.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614–622.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 254–263.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 70–74. Association for Computational Linguistics.

Tran-Thanh, L., Huynh, T. D., Rosenfeld, A., Ramchurn, S. D., & Jennings, N. R. (2014). BudgetFix: Budget Limited Crowdsourcing for Interdependent Task Allocation with Quality Guarantees. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pp. 477–484.

Tran-Thanh, L., Venanzi, M., Rogers, A., & Jennings, N. R. (2013). Efficient Budget Allocation with Accuracy Guarantees for Crowdsourcing Classification Tasks. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, pp. 6–10.

Venanzi, M., Guiver, J., Kohli, P., & Jennings, N. R. (2016). Time-sensitive bayesian information aggregation for crowdsourcing systems. *Journal of Artificial Intelligence Research*, *56*, 517–545.

Welinder, P., & Perona, P. (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pp. 25–32.

Yan, Y., Rosales, R., Fung, G., & Dy, J. G. (2011). Active Learning from Crowds. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1161–1168.

Zhang, Q., Wen, Y., Tian, X., Gan, X., & Wang, X. (2015). Incentivize crowd labeling under budget constraint. In *Proceedings of 2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2812–2820.

Zhu, C., Xu, H., & Yan, S. (2015). Online Crowdsourcing. *arXiv:1512.02393 [cs]*.

Zhuang, H., & Young, J. (2015). Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM ’15, pp. 243–252.