Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error Compared to the Majority-vote Method

HASTI NARIMANZADEH, Aalto University, Finland

ARASH BADIE-MODIRI, Aalto University, Finland and Central European University, Austria

IULIIA G. SMIRNOVA, University of Helsinki, Finland

TED HSUAN YUN CHEN, George Mason University, USA

How to better reduce measurement variability and bias introduced by subjectivity in crowdsourced labelling remains an open question. We introduce a theoretical framework for understanding how random error and measurement bias enter into crowdsourced annotations of subjective constructs. We then propose a pipeline that combines pairwise comparison labelling with Elo scoring, and demonstrate that it outperforms the ubiquitous majority-voting method in reducing both types of measurement error. To assess the performance of the labelling approaches, we constructed an agent-based model of crowdsourced labelling that lets us introduce different types of subjectivity into the tasks. We find that under most conditions with task subjectivity, the comparison approach produced higher f_1 scores. Further, the comparison approach is less susceptible to inflating bias, which majority voting tends to do. To facilitate applications, we show with simulated and real-world data that the number of required random comparisons for the same classification accuracy scales log-linearly $O(N \log N)$ with the number of labelled items. We also implemented the Elo system as an open-source Python package.

CCS Concepts: • **Information systems** → **Crowdsourcing**.

Additional Key Words and Phrases: crowdsourcing, subjectivity, comparison method, majority-vote method

1 INTRODUCTION

Human labels are consistently seen as the gold standard for producing data used to train or validate models and systems. These tasks include assessing image characteristics [46, 61], labelling text generated by humans or machines [2, 32, 38, 39], validating images produced by generative adversarial networks [6, 13, 47], doing last-mile validation of AI predictions [41, 42], and creating specifications for data set evaluation [5]. Due to demands for increasingly larger data sets, a substantial portion of such subjective manual labelling tasks has shifted toward crowd work in recent years [33, 45]. However, despite their ubiquity in data collection and measurement tasks,

Authors' addresses: Hasti Narimanzadeh, hasti.narimanzadeh@aalto.fi, Aalto University, P.O. Box 11000, Finland, FI-00076; Arash Badie-Modiri, arash.badiemodiri@aalto.fi, Aalto University, P.O. Box 11000, Finland, FI-00076 and Central European University, Quellenstraße 51, Austria, A-1100; Iuliia G. Smirnova, julia.98.g.s@gmail.com, University of Helsinki, P.O. Box 4 (Yliopistonkatu 3), Finland, FI-00014; Ted Hsuan Yun Chen, ted.hsuanyun.chen@gmail.com, George Mason University, 4400 University Drive, Fairfax, Virginia, USA, 22030.

extracting and aggregating subjective information from crowdsourced human coders remain open areas of research [58, 64].

In this paper we focus on a particular class of subjective crowdsourcing tasks where the target quantity is a *subjective construct* characterised by inherent subjectivity [51], such as content toxicity [2, 5, 50], as opposed to one where there can be objectively defined targets [30, 55]. *How do we obtain a valid classification or rating on a set of items for some subjective construct from a group of subjective human coders?* Consider, for example, the problem of identifying or quantifying the level of toxicity in online content [2, 5, 50]. The to-be-fit models for toxicity detection work by being optimised to best predict the average coder's assessment of toxicity, making input data critical to their performance. It is therefore integral to recognise that human assessment of toxicity is at best intersubjective and will be affected by numerous "hidden variables", namely social, linguistic, and economic factors that are often not explicitly observable in the assessment. As we describe below, these unobserved subjective factors increase random error and can often also introduce factors that researchers reasonably wish to disassociate from the construct in question (e.g., for algorithmic fairness purposes), which can be understood as a type of measurement bias [7, 28, 37]. Crowdsourced toxicity labels, for example, have unfortunately been found to be positively associated with terms related to LGBTQ and race topics [15].

Prior efforts have been made to improve labelling through subjectivity-aware tasks, including enhancement of questionnaire and task design [28, 62], an example being the quality-control workflow by Drapeau et al. [16], which asks crowd raters to justify their answers or re-evaluate their decisions when faced with counter-arguments; uncovering and utilising disagreement and their sources among raters [9, 31, 49]; model- or aggregation-based bias correction methods that rely on rater task histories [35, 37, 57, 60]; and devising methods to collect labels from unskilled crowd raters that are on par with or superior to those from raters with domain expertise [18, 27, 52]. Just as important, others have made efforts to conceptually and empirically sort through the nature of subjectivity in human-labelling tasks, such as classifying tasks by their level of subjectivity [51] or identifying the different sources of subjectivity [17, 31].

We add to this body of work by proposing a combined task design and aggregation approach for crowdsourcing measures of subjective constructs. Specifically, we describe a pairwise comparison-based labelling task, shown in the top row of Figure 1, where raters are given tasks of comparing the strength of the construct in question between two items, which we aggregate using the Elo rating system. We compare the pairwise comparison approach to the classic "majority-vote" labelling method, a ubiquitous mechanism where the same item is assessed by a specific number of people and the label for each item is selected based on the decision of a majority of voters. We are not the first to propose relative assessments for crowdsourced labelling [10, 11, 44, 56, 65], which we summarise below, but we do so explicitly in the context of obtaining subjective constructs, and combine the comparisons with a straightforward aggregation method, an open-source implementation of which is provided with this paper.

Further, we contribute to the definitional clarity of "subjectivity" by presenting a conceptual framework for crowdsourcing subjective constructs that maps this type of crowdsourcing task onto concepts from survey methodology, an idea that has been proposed elsewhere [1]. This framework lets us formalise what "subjectivity" means in measurement tasks, and show how subjectivity increases random error and measurement bias in aggregated labels. Namely, subjectivity increases variance and therefore increases random error, and coupled with necessarily imperfect task definitions means there are likely going to be sources of bias. We show that our pairwise comparison method reduces random error, as measured by binary-label f_1 scores. At the same time, while our method does not completely mitigate the issue of bias (which should be addressed in

other steps of the rating pipeline [37]), it outperforms the majority-vote method which actually inflates the inherent bias in the system.

Our work contributes to a wide range of literature that engages with the question of crowdsourcing and how to measure subjective constructs [17, 28, 31, 51, 60, 62]. In addition to addressing the topic at a theoretical-level, mapping the task to a survey methodology problem, our work has clear implications for applied research. To facilitate applied work, we provide a reference implementation of the Elo rating system for pairwise comparisons as an open-source Python package under the MIT license.[1]

Our paper proceeds as follows. Beginning with Sec. 2, we define our task of obtaining a valid subjective construct and formally describe the behavioural model of crowdsourced labelling. Here, we discuss how subjectivity enters the model, and how it introduces measurement error. In Sec. 3, we draw on previous work about comparisons and voting to discuss the advantages of comparison approaches over voting-based ones, both in terms of construct validity and implementation cost. Next, in Sec. 4, we describe our proposed approach and our experimental setup. In Sec. 5, we present our results. We show that (1) for a wide range of subjective problems, the comparison method performs better than majority voting and the gap often gets wider the more subjective a task becomes; (2) the comparison method is more resilient towards spam, compared to the majority-vote method; and (3) the number of required comparison grows log-linearly with the number of items when used with the Elo rating system, not quadratically.

2 CROWDSOURCING SUBJECTIVE CONSTRUCTS

Building on previous research, we understand crowdsourcing measures for subjective constructs as a descriptive inference task of estimating the distribution of *perceptions toward how items relate to a researcher-defined construct*. Such answer distributions can more accurately represent how people perceive the construct under study, and so the quest for devising crowdsourcing tasks conducive to estimating the population distribution as accurately as possible is an active area of research [12, 19].

Exactly which aspects of the distribution to use in downstream research is an open question that touches on the existence and the nature of ground truths in subjective constructs [3, 64]. A longstanding debate in the legal community surrounding how to define subjective constructs is instructive here. Namely, defining "obscenity" has historically been a challenge for the U.S. Courts system. It is a highly subjective term with immense legal implications [24]. Going back as far as *Roth v. United States* in 1957, the U.S. Supreme Court has constructed the hypothetical judge of obscenity to be the "average person, [who is] applying contemporary community standards". U.S. Supreme Court Justice Potter Stewart's famous 1964 statement "I know it when I see it" about

---

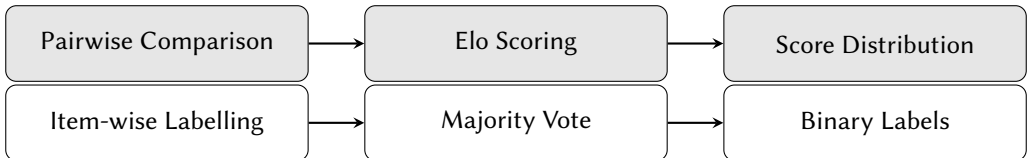[1]Available at https://github.com/hastinarimanzadeh/elo-rating.

Fig. 1. Different labelling approaches to obtaining measures of subjective constructs. The top row, highlighted in gray, shows our proposed pairwise comparison approach, which yields a continuous distribution of scores (e.g. level of toxicity). The bottom row shows a traditional approach based on binary item-wise labels (e.g. toxic or non-toxic).

defining "hard-core pornography" was in reference to *Roth*, and was used then, and elsewhere [53], to justify a data-driven and inductive way of defining subjective constructs. Indeed, when working with purely subjective constructs, as opposed to extraction of factual information [55] or skill-based assessments [30], the crowd from which we derive the average judgement serves not as aggregated intelligence but instead as a reasonable *definition* of the inter-subjective ground truth.

With this in mind, we proceed in this study with the premise that the population distribution surrounding perceptions of how items relate to the subjective construct of interest is usefully modelled as a unimodal distribution whose mean is meaningful. Specifically, we will treat the population mean as the "ground truth" and the target of the crowdsourcing task.[2] This problem definition sets our study apart from others that focus on using crowd intelligence to complete tasks whose outcomes can be evaluated objectively [e.g. 30, 55].

Following this logic, crowdsourcing subjective constructs and the challenges associated with this can be understood in terms of concepts from survey methodology. Specifically, our target is the population mean and we sample individuals from the population to estimate this quantity. Measurement error (i.e., deviations between the quantity inferred from the sample and the target measure) is either due to random error from sampling variation – which is amplified in crowd-sourcing tasks since usually only a very small number of responses are obtained – or a type of response bias stemming from mismatch between rater responses and the target construct, which includes those due to under-defined tasks or unfair perceptions held by the crowd raters.[3]

"Subjectivity", broadly defined as variations in how constructs are understood, fits into this framework as it contributes to measurement error by increasing the population variance surrounding the construct, thereby increasing random error, or by introducing unwanted components to the construct, thereby adding response bias. We discuss these in detail below.

2.1 Model of crowdsourced labelling and sources of subjectivity

In subjective crowdsourced tasks, definite answers can be elusive, not to mention susceptible to sources of inter-rater disagreement. Linguistic philosophers have long recognised different sources of ambiguity in meaning stemming from what corresponds in the data-labelling context to the ambiguity of data, differences in raters' perspectives, and ambiguity in annotation labels [34, 43]. These concepts have recently been adapted to research on subjectivity in crowdsourcing [3, 17, 31]. Kairam and Heer [31], for example, identified various sources of disagreement among sub-groups of crowd raters in their entity annotation tasks of Twitter and Wikipedia data. Drawing from these concepts, we mathematically formalise our conception of subjectivity in crowdsourcing tasks below.

Following the previous discussion, we devised a simplified mathematical model of human perception and labelling of subjective constructs. We begin with an observable item, such as an internet comment, and a human rater, whom we ask to rate, for example, the item's level of toxicity. Intuitively, the rater will subconsciously (1) observe all aspects of the item, (2) give a toxicity component score to each observable aspect, (3) sum the component scores for a total item toxicity score, and (4) assign the item a toxicity label or numeric score, or compare it to another item.

2.1.1 Definitional subjectivity. Let us encode all the objective observable facts about the item (e.g. number of times profanity was used) as a very large vector of numbers $\vec{X}$. The elements of this
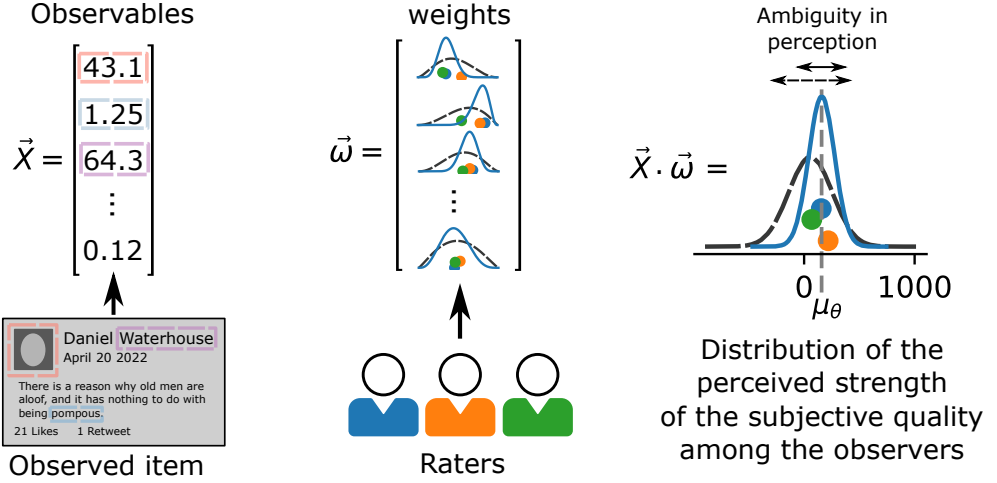

Fig. 2. A simplified mathematical model of human understanding of subjective constructs. Let us assume vector $\vec{X}$ encodes all observable features of an item. For each rater there exists a vector of weights $\vec{\omega}$, normalised so that $\|\vec{\omega}\| = 1$, where ω_k determines the contribution of the observable X_k in the strength of the subjective construct perceived by that rater θ_i. The distribution of weights ω_k in the population of raters determines the overall ambiguity of perception for that item in the population. A more diverse set of opinions in the population, visualised here as the more widely distributed weights with dotted lines, produces more perceived strengths that are more varied across the population of raters. Conversely, more agreement in the weights, shown using solid blue lines, will result in higher overall agreement. In this paper, the standard deviation of the distribution of perceived strengths of the subjective construct is denoted as *ambiguity in perception*.

vector independently encode all available information about the item.[4] In this simple model we assume that each element of $\vec{X}$'s contribution to the subjective construct is weighed by the rater according to a vector of weights $\vec{\omega}$, so that

$$\theta = \vec{X} \cdot \vec{\omega} = \sum_k X_k \omega_k \,, \tag{1}$$

with the vector of weights being (arbitrarily) of unit length $\sum_k \omega_k^2 = 1$. The vector $\vec{\omega}$ is specific to each rater and is affected by their personal idiosyncratic preferences. Essentially, $\vec{\omega}$ is the rater's definitional understanding of the construct, and subjectivity exists because it differs across raters. The resulting value of θ denotes a rater's perception of the strength of the subjective construct in question. Figure 2 shows a schematic visualisation of this model. As defined above, our goal is to devise a method of estimation that recovers μ_θ, i.e., the population mean perceived strength of the subjective construct.[5]

We define subjectivity as *variation across individuals' conceptualisation of a given construct* (i.e., understanding and judgement) based on individual idiosyncrasies. If there is any variation in the population distribution of this conceptualisation that is not attributable to factors external to these individuals, we say there is subjectivity surrounding the construct and therefore the rated item.

---

[4]Dependence between elements a and b can be account for by adding in an additional element capturing the interaction term $a \times b$.

[5]It is not possible to know without extensive study the distribution of ω_k in the population. We can, however, reasonably assume, based on generalisations of the Central Limit Theorem, that for an item with a large enough number of observables, the value θ_i of the same item across the rater population is sampled from a normal distribution $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$.

Subjectivity factors into labelling tasks in multiple ways. At the definitional stage, which we just described, variations in preferences and experiences across individuals influence how much weight ($\vec{\omega}$) different things (i.e., observables in $\vec{X}$) are given when measuring a construct [2, 28]. *Does profanity matter at all to text being perceived as toxic? If so, how much?* This kind of variation in what "counts" is a major source of subjectivity in labelling tasks. This type of subjectivity can be mitigated by clear coding rules [2] – *do not consider the speaker's race when assessing the toxicity of what they said* – but given the effectively infinite feature space of plausible observables in $\vec{X}$, subjectivity cannot be eliminated. Further, because many of these internal assessments are deeply-rooted biases, explicit prompts about biases can still fail to mitigate them [28].

As subjectivity about a given construct increases, the population variance of each ω_k increases, which is then reflected in increased variance in the distribution of θ. The standard deviation σ_θ therefore can be understood as the *ambiguity in perception*, which stems from the definitional process surrounding a subjective construct. Larger values of ambiguity in perception imply there will be higher variation in the perceived strength of the subjective construct among the population for a given item. As we illustrate in Fig. 2, the consequence of higher ambiguity in perception is that there will be higher measurement error in the form of random error stemming from sampling variation, as the greater the population variance, the more likely a given sample mean will fall from the population mean. Standard crowdsourcing practices exacerbate this type of measurement error because samples (i.e. the number of raters per item) tend to be small.

2.1.2 Translational subjectivity. Another source of subjectivity comes from raters' mental model when translating the latent θ into a concretely expressed measure, which can be through a classification task – *this comment is toxic* – or a comparison – *comment A is less toxic than comment B*. Specifically, θ_i exists internally for each individual, and the act of encoding it (e.g. threshold-based classification, quantitative rating, comparison) is also subject to variations across individuals.

If the rater is asked to make a binary classification, they compare their perceived strength of the subjective construct θ_i to some personal threshold $\theta_i^{\text{threshold}}$. As with the weight vector $\vec{\omega}$, the threshold is also specific to each rater and parameterise the sensitivity of the rater to the subjective construct. We refer to this type of subjectivity as *diversity in threshold*.

On the other hand, if the rater is asked to make a comparative assessment of two items, that is, to determine which of two items is stronger on the subjective construct, the rater will perceive the magnitude of the construct in the two items, say θ_1 and θ_2, and then compare them. Based on prior work on rater agreement, it is reasonable to assume that it is easier for a rater to compare two items that are perceived to be more different to each other with respect to the strength of the subjective construct in question [48], i.e., when $|\theta_1 - \theta_2|$ is larger. In our model, the rater is not able to make a reliable comparison if the difference in perceived strength of the two items $|\theta_1 - \theta_2|$ is smaller than some cutoff value δ_θ. We refer to this parameter δ_θ as the *ambiguity in comparison*.

2.1.3 Subjectivity-induced bias. Finally, we discuss the special case where definitional subjectivity, discussed above, becomes a source of measurement bias. Specifically, instead of targeting the population mean μ_θ (i.e. where the researcher recognises the existence of definitional subjectivity but is agnostic about what "counts" in the construct's definition), there are instances where researchers would wish to exclude certain features from the definition of a construct and therefore not have them factor into the crowdsourced measure. In these cases, if researchers do not or cannot fully define the task, which we established is difficult [28], the presence of subjectivity among the raters about whether these features should be a part of the construct's definition means μ_θ becomes a biased version of the target construct, which we illustrate in Fig. 3 with a schematic visualisation.

Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error
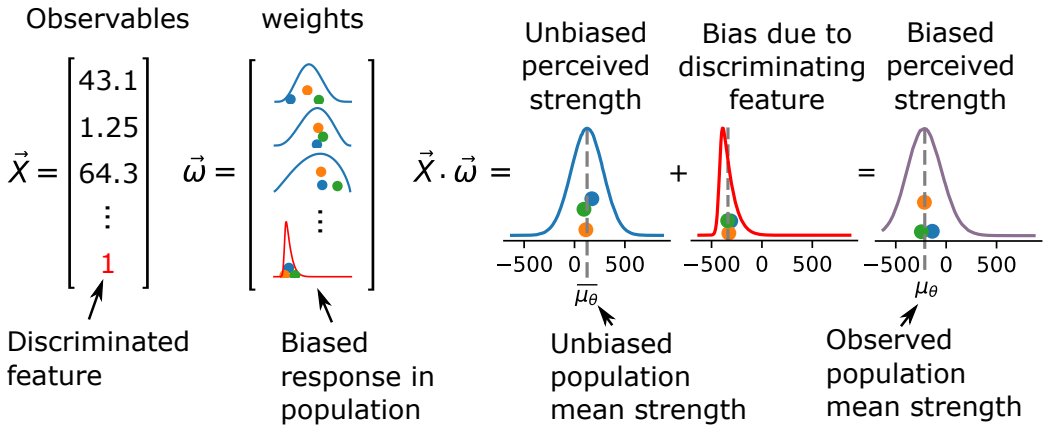Compared to the Majority-vote Method

334:7

Fig. 3. A simple mathematical model of bias induced by the existence of a discriminatory feature in an item. Some items might display features that the researcher decides, for ethical or scientific reasons, should not affect the perceived strength of the subjective construct. Existence of LGBTQ and race-based terms contributing to toxicity labels exemplifies this issue [15]. In the simplified model of human understanding of subjective constructs presented here, these features induce an error (and no bias) if the distribution of their associated weights has a mean of zero. On the other hand, if the discriminated feature has a weight distribution with a mean much larger or smaller than zero, it induces a bias in the perceived strength of the subjective construct in all items that exhibit the feature. Here, we see that a heavily-skewed distribution of weights (in red) associated with the discriminated feature significantly affects the perceived strength of the subjective construct for the items exhibiting that feature.

In our model, this scenario is presented as researchers desiring an element in $\vec{\omega}$ to be zero when its population mean is in fact nonzero.

What we have described here is effectively a form of response bias stemming from the inability of researchers to control unwanted definitional subjectivity. This scenario is becoming increasingly common as researchers start to recognise the importance of excluding unfair or otherwise undesirable observables from entering their labelled data, and in turn their models or algorithms. For example, researchers classifying internet comments into toxic and non-toxic categories are likely to want the labels to be unaffected by ethnic or racial features present in the profile image of the poster.

3 COMPARISON VERSUS MAJORITY-VOTE METHODS

Having now defined our target (i.e. a subjective construct measured as its population mean μ_θ) and presented the associated challenges stemming from different sources of subjectivity, we move to discuss the relative benefits of two widely used ways to formulate crowdsourcing tasks, the comparison and the majority-vote methods which we presented in Fig. 1. Drawing from previous literature, we discuss how the two methods compare in terms of the validity of the obtained measure when they are faced with different sources of subjectivity. We also review claims about their costs.

3.1 Validity

The two methods we study in this paper have been previously compared elsewhere in the literature. One line of research focuses on the cognitive benefits of pairwise comparison tasks over item-wise

labelling tasks at the individual-level, while another considers how these tasks compare specifically in crowdsourcing contexts where information gathering tends to favour breadth over depth.

3.1.1   Cognition-based benefits of comparative over absolute judgements. From the psychology and cognition literature, Goffin and Olson [23] argue that when individuals assess performances, comparative judgements are superior to absolute judgements, which are the building block of the majority-vote method, because the latter often lacks a clear anchor. On the other hand, anchoring assessments to an external comparison point provides respondents "shared reference points on which to base their judgements", which "may reduce error variance by increasing respondents' agreement about the meaning of scale points [23, p.53]". Gentner and Markman [22] find a similar mechanism, that explicit juxtapositions highlight similarities and differences, which provides clarity on the compared construct. This concept has been usefully applied to human-labelling contexts, such as using comparison-based crowdsourcing to extract features [11, 65].

3.1.2   Removing diversity in threshold through relative anchoring. Other work from human-labelling fields, such as image assessment [44] and toxicity labelling [2], focus explicitly on the benefits of *relative anchoring*, where to-be-labelled items are assessed against each other, dispensing with the need for either individuals' internal anchors or external anchors. In their seminal piece on relative assessments for image labelling, Parikh and Grauman [44] offer assessing smiling as an illustrative example of the relative anchoring concept. Given three images of (1) someone clearly smiling, (2) someone clearly not smiling, and (3) someone having a borderline expression between smiling and non-smiling, it is difficult for absolute ratings to place the third category, but the comparative rating can always confidently distinguish between any pair of images.

This argument about relative anchoring directly maps onto our discussion of translational subjectivity from Sec. 2.1.2. Specifically, because compared items serve as reference points to each other, the comparison method is able to remove diversity between raters' thresholds, one of our identified sources of subjectivity, from the labelling procedure.

3.1.3   Avoiding measurement error inflation for high-subjectivity constructs. Another line of work focuses on how the majority-vote method obscures the amount of uncertainty surrounding specific items [2, 49]. By construction, the majority-vote uses popularity to adjudicate between two side of a binary label, thereby forcing agreement. However, the very fact that the majority-vote needs to be employed in the first place implies there is variation in the population surrounding the subjective construct being measured, which can be shown through inter-rater agreement scores [1, 49, 50].

The discarded labels are not random, but are instead due to the raters' subjectivity, which as we argued can stem from both definitional and translational reasons. Further, the discarded data will always be from minority raters holding less popular opinions. The majority-vote method, then, moves the estimated mean $\hat{\mu}_\theta$ further from the true population mean by non-randomly discarding up to $[\frac{1}{3}, \frac{1}{2})$ of the ratings. As measurement error due to population variability in crowdsourcing tasks is already likely to be high because only a small sample size is used for each item, this kind of data removal will further exacerbate the problem. Taken to the extreme, the majority-vote method is equivalent to running a survey and claiming that the population can be characterised by the modal response category.

Instead, as Aroyo et al. [2] argue, for inherently subjective tasks (e.g., measuring subjective constructs as we do), a comparative ranking would result in better overall agreement between raters compared to an item-wise absolute rating due to the fact that the comparison method can capture the diversity of opinions between raters as opposed to optimising for consensus. This is achieved when using a pairwise comparison approach, as it does not discard any data. In fact, more than not wasting data, because the comparison approach allows each item to be observed multiple

times but as part of different unique item pairs, it provides more information for aggregation algorithms such as the Elo method we describe in Sec. 2 to produce a more accurate estimate of θ's population distribution.

3.1.4   Avoiding bias and spam inflation. Finally, following the immediately preceding discussion from Sec. 3.1.3, we hypothesise that comparisons will outperform majority-voting over item-wise labels when dealing with the kind of subjectivity-induced bias we discussed in Sec. 2.1.3. We start by recognising that should this type of bias exist, it will affect both methods. By Eq. (1), definitional subjectivity (captured by the presence of unwanted elements with nonzero mean in $\vec{\omega}$) is a part of θ, which means that because both the comparison and majority-vote methods target the population mean μ_θ, neither method do anything to *remove* the bias.

However, as the majority-vote discards observations based on popularity, it can yield highly biased estimates for some items due to random sampling, which we speculate will *inflate* the extant population bias.[6] On the other hand, because pairwise comparisons are able to more accurately reflect the population distribution of θ, it should also more accurately capture the unwanted bias component in μ_θ.

The superiority of the comparison method here, then, is in being able to avoid inflating measurement bias. A similar line of reasoning regarding the aggregation of unwanted information can be applied to the effect of spammers (i.e., raters who complete tasks randomly or semi-randomly), which is an active research area in crowdsourcing [14, 59].

3.2 Cost

Despite all the ostensible benefits of using comparisons, costs, specifically the cost of performing all possible number of pairwise comparisons grows quadratically with the number of items being rated. This serves to be an impediment to adopting the comparison method over the majority-vote method, whose total possible number of tasks grows linearly with the number of items, especially when dealing with large data sets [25]. Further, previous studies have hypothesised that a certain level of labelling quality requires many more comparison tasks or an algorithmically selected set of comparison tasks, compared to the majority-vote tasks [25, 30]. Ye and Doermann [63], for instance, propose a method of augmenting the comparison method with absolute judgement to improve the labelling quality. They also propose an active method for determining the next batch of comparisons and absolute judgement tests based on previous batches of results. Similarly, Jang et al. [30] and Guo et al. [25] propose an active method for determining future comparisons to reduce the number of comparisons required to arrive at a full ordinal ranking.

In this paper, we show that depending on the subjectivity of the problem a comparable number of comparison tasks can provide as good, or better quality of labels compared to the majority-vote tasks, and that for a randomly selected set of comparisons combined with a simple implementation of the Elo rating system, the number of required comparisons for a constant level of labelling quality scales log-linearly, not quadratically, with the number of items.

Finally, we note here that researchers have argued that comparisons are often easier and faster to make compared to the majority-vote method when assessing the relevance of search results in information retrieval systems using crowd-sourcing [8]. A faster and easier task very commonly means that a higher number pairwise comparison tasks can be performed with the same amount of time and effort spent by the raters compared to tasks that rely on an absolute single-item judgement.

---

[6]Depending on the distribution of weights ω_k for the bias-inducing element in $\vec{\omega}$, this behaviour could plausibly reduce bias, but this is largely idiosyncratic to particular cases and therefore cannot be relied on as a feature of the method.

4 METHODS

In the remainder of this paper, we explore questions raised in the preceding discussion regarding the relative strengths of the majority-vote and pairwise comparison methods when crowdsourcing measures for subjective constructs. We compare each method's response to the different types of subjectivity we discussed in our theoretical framework, including both definitional and translational sources of subjectivity. We also assess their performance in dealing with the presence of subjectivity-induced bias as we hypothesised in Sec. 3.1.4. Finally, we study the scaling behaviour of the pairwise comparison method to establish how the number of required comparisons to keep the same level of accuracy changes as the number of items to be rated grows.

Before moving to our results, we present the methods we used. First, in Sec. 4.1, we introduce the label aggregation approach we adopt in this paper, namely the Elo rating system. Next, we describe in Sec. 4.2 how we quantify our subjectivity formalisation through the simulation of items and raters. Finally, in Sec. 4.3, we summarise our experimental pilot of Twitter conversations which we use to corroborate our simulated $O(N \log N)$ scaling of random comparisons.

The code required for reproducing the numerical results presented in this paper, along with all the necessary instructions and data, are available as a Git repository at https://github.com/hastinarimanzadeh/elo-paper-reproduction.

4.1 The Elo rating system

As we showed in Fig. 1, the results from pairwise comparisons undergo an algorithmic transformation before ending up as ratings, which can then be binarized by rank for classification purposes. The problem of assigning ratings to the strength of each player based on their performance in one-on-one matches against others appears frequently in games and sports. Historically, hand-crafted rating systems relied on an often intrinsic judgement of each player's merit or achievements in each game, combined with a myriad of statistical and heuristic methods specific to the game in question. In this paper, we dispenses with these historical approaches and propose using the Elo rating system as a simple yet effective method of aggregating the results of comparisons performed by crowdsourcing raters.

The Elo system works by relying on statistical estimation of ratings solely based on the final outcome of performed matches. In short, for each match, the Elo system makes a prediction of the outcome probabilities in a zero-sum games based on the two participating players' difference in rating. The two players *wager* some fixed amount k of their rating based on the predicted probabilities. If the outcome of the match is closer to the predicted outcome (e.g., a much stronger player wins against a much weaker player) fewer rating points change hands. In the case of an unlikely outcome, a larger share of wagered points is transferred to the winning player.

The predicted outcome of a match depends on the difference in the ratings of the two players. The system assumes that the actual game performance of a player in a single game is a random variable, often assumed to be normally or logistically distributed, with a mean equal to the player's true rating (representing their aptitude or strength) and a scale arbitrarily set. The Elo system makes the simplifying assumption that all players have the same scale for their *game performance distribution*. A single match between two players can be simulated by each player drawing a number from their respective game performance distribution. The probability of a positive outcome for each player is equal to the probability of their drawn number having a larger value.

Many real-world rankings, such as ranking of countries by population and performance of sports players, show a heavy-tail decay of scores as ranks increase [29]. At first glance, it might seem that the Elo system assumes a specific distribution for the game performance. Instead, the system makes an assumption about the *distribution of relative performance of two players* (i.e. their respective

Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error
Compared to the Majority-vote Method

334:11

probability of winning) given their difference in rating. In other words, the inherent assumption of this method is of a normal performance distribution, often called the Thurstone–Mosteller model, or based on a logistic performance distribution referred to as Bradley–Terry model. Though it has been shown that in a practical setting the two models produce virtually equivalent results [54] the latter model, based on logistic distribution, is commonly used for ranking human competitive behaviour such as rankings for chess players and in online multi-player games, possibly due to its nicer analytical behaviour [40].

A commonly used formulation, based on the logistic distribution assumption is as follows:

$$E_A = \frac{1}{1 + 10^{\frac{R_A - R_B}{400}}}, \tag{2}$$

where R_A and R_B indicates ratings (aptitude) of the players A and B. E_A indicates the probability of a positive outcome of the game for player A. As possible outcomes range from 0 to 1 the expected outcome for each player is the same as the probability of a positive outcome for that player. The use of base 10 and denominator 400 makes it easier to mentally compare rating values, as a difference of 400 rating units is equivalent to a ten times larger expected score.

After a single game, the ratings are adjusted based on the difference in the realised outcome compared to the expected outcome of the game

$$R'_A = R_A + k(S_A - E_A), \tag{3}$$

where S_A is the realised outcome of the match for player A. The rating for player B is also adjusted using the same formula. The parameter k is a tuning parameter, roughly similar to the learning rate parameter in many machine learning methods, which indicates the maximum possible point adjustment per match. A high value for k results in larger adjustments to the ratings after each match, at the cost of higher overall noise. Conversely, selecting a lower value of k requires more matches to arrive at equilibrium in ratings.

In addition, we found it often helpful in practice to run the rating system with multiple cycles on the available set of comparisons, generally similar to the deep learning common practice of cycling learning data in multiple epochs. At each epoch, the set of all available comparisons are used in random order for updating the ratings. This can be repeated until an equilibrium is reached.

The Elo rating system has nice properties for our application. While the zero-sum nature of the Elo rating system might at first glance appear a weakness, we argue that this fits well with our goal of labelling items based on the population average. As results of each comparison changes the rating of two items by equal amounts but in opposing directions, the mean ratings value of all items will always stay the same. If all items start with an initial rating of zero, the label of each item can be determined by simply checking the sign of its estimated rating. The rating is directly related to the probability of that item being selected in a comparison with a hypothetical item that possesses the exact average of the intended subjective construct compared to all other items as construed by the population of the raters, as can be surmised from Eq. (2).

We also elected to use the Elo rating system in part due to its simplicity. In the next two sections we will show both empirically and using a simple mathematical model that despite relying on the most simple approach of randomly comparing items and aggregating the results using the Elo rating system, we still arrive at a scalable and robust estimation of popular average labelling of subjective constructs.

An open-source reference implementation of the Elo rating system is provided as a Python package in the official Python Package Index. The package can be installed using the command `pip install elo-rating` with the documentation available at https://github.com/hastinarimanzadeh/elo-rating. The package provides an Elo class that computes ratings and rankings based on a

list of provided results of comparisons. Further details on the interface are delineated in the documentation.

4.2 Simulation

The relative performance of the comparison and majority-vote methods when faced with different sources of subjectivity can be studied using a simple agent-based model of the problem. We present here such a model, which encompasses the subjectivity model we previously described in Sec. 2.1 and schematically illustrated in Fig. 2. We follow this with details on the comparison and majority-vote methods simulation.

4.2.1 Setup. Let us assume N items, e.g., internet comments, each with a "true" rating, indicating the inherent value of the construct in question as it would be judged by the average rater for each of the items, e.g., toxicity. For the purposes of this simulation, we draw the true ratings from a normal distribution with an arbitrary mean of 0 and a standard deviation of 1. A perfect binary classification system would be able to distinguish all items with ratings above 0.

We also simulate M raters, each with a personal threshold of what they would consider a significant enough presence of the intended construct, e.g., strong enough toxicity that impels them to recognise a certain comment as toxic. The personal thresholds are drawn from a normal distribution with the same mean as the mean true rating of items and standard deviation set by the parameter *diversity in threshold*, directly mapping to the translational subjectivity we discussed in Sec. 2.1.2.

To contextualise the definitional subjectivity we discussed in Sec. 2.1.1, for each item a rater observes a perceived rating for that item, drawn from a normal distribution with mean set to the true rating of the item and a standard deviation set by the parameter *perception ambiguity*. The perceived rating of one item for one rater is fixed across different observations by the same rater at least in the time scale of one study, as it is assumed to be a function of past experiences, internal priorities, and their definitions of the construct in question. A rater can classify one item by comparing their perceived rating of that item with their *personal threshold*.

Table 1. The set of parameters used in the agent-based simulation setup.

Parameter name	Description
Ambiguity in perception	Standard deviation of raters' perceptions θ, defined in Sec. 2.1.1
Ambiguity in comparison	Minimum difference between perceived strength of two items ($|\theta_1 - \theta_2|$) below which the two items cannot be ordered by a rater, defined in Sec. 2.1.2
Diversity in threshold	Standard deviation of raters' personal thresholds for the subjective construct, discussed in Sec. 2.1.2
Rater bias	Rater's weight for a discriminatory feature in the observables vector, defined in Sec. 2.1.3
Fraction of spammers	Fraction of spammers, raters who vote or compare randomly
Number of items	Items to be labelled
Number of comparison tasks	Pairwise comparisons among all items
Number of votes per item in majority-vote	Unless stated otherwise, 3 votes per item

A rater can also be given a comparison task, for which they compare two items based on the perceived ratings of the items. It is, however, inherently more difficult to confidently compare items when the difference in the magnitude of the intended construct is small. We identified such an ambiguity to stem from translational subjectivity in Sec. 2.1.2. To allow for this in our model, if the difference in the perceived ratings of the two items is smaller than a certain threshold, as determined by the parameter *comparison ambiguity*, the rater identifies them as equal. Conversely, a *comparison ambiguity* of zero results in no rater selecting the equal option, but simply selecting whichever item that corresponds to a higher perceived rating for that rater.

Spam is an unfortunate reality in crowd-sourcing. For some problems, it is possible to systematically detect spammers based on various methods. Kuang et al. [36] for instance uses a heterogeneous network embedding approach to detecting spam raters based on defined collusion patterns. These methods become more difficult to implement for the case of subjective problems [1]. To assess the susceptibility of the two methods to spam, a fraction of raters can be simulated as spammers, essentially voting or comparing presented items at random regardless of the items' ratings or perceived ratings.

4.2.2 Simulation. For each item, three random raters are drawn from the pool of M raters (which might include spammers) and each rater is asked to provide a binary classification of that item (a vote) based on comparing their perceived rating of that item and their personal threshold. The three votes are recorded and items are labelled using a majority-vote system. Next, a number of pairs of items are drawn from the pool of all possible combinations of size 2 of N items for comparison. Each item pair is passed to a rater who compares them based on their perceived ratings and comparison ambiguity. A binary classification is established by running the Elo ranking algorithm on the comparison results and classifying items based on their final Elo ratings being larger than 0.

4.2.3 Evaluation. For each combination of parameters, the performances of the two classification methods are measured as compared to the ground-truth binary labels based on the true rating of each item. For both methods, the positive label f_1 score of each run is calculated. The mean difference in f_1 scores across an ensemble of 25 independent runs determines which binary classification method worked best in each specific scenario.

4.3 Scaling experiment: Controversiality of climate change conversations on Twitter

To confirm the scaling behaviour and the properties of the comparison method, we also performed a set of real-world experiments. In these experiments, the crowdsourced task was to classify a conversation on Twitter, consisting of the main tweet and its whole response tree, as either controversial or non-controversial. The goal of the experiment is to study the scaling behaviour of the comparison method and provide an example of how a small-scale pilot study can help estimate the number of required comparisons for a larger study.[7]

To this end, we selected a sample of 41 conversations related to the topic of climate change. We sampled these conversations from all Twitter conversations between February 2019 and September 2021, where the original post contains at least one of the list of 13 keywords or 36 hashtags commonly

---

[7]As discussed in our problem definition (i.e. Sec. 2), when working with subjective constructs like the controversiality of a conversation, the "ground truth" for a given item is that item's population mean perception μ_θ. This target quantity can only be estimated, as even when working with a population census, which removes random error, labelling is still subject to the different types of subjectivity we outlined above. While it is possible to conduct a large-scale empirical study aimed at extracting the population distribution of each conversation as precisely as possible for the purpose of benchmarking different crowdsourced labelling approaches, we believe this is beyond the scale of our current study. We therefore focus our real-world experiment on studying scaling properties.

associated with climate science, climate action, or climate scepticism, and the conversation involves at least three unique users.[8]

We constructed a set of crowdsourcing tasks for all 820 possible comparisons between two given conversations. The comparison task stated the following prompt: "Read the following two Twitter conversations. Which of the following two conversations shows a *higher level of disagreement* among its participants?". Each time a conversation was presented in a task, the replies were presented in random order and limited to a maximum of five first-level replies and their second-level replies. Each of the 820 possible comparisons was presented to two raters on Amazon Mechanical Turk. The raters had three possible choices for each comparison. They could select the conversation on the left, on the right, or judge them to have an equal level of disagreement. Each comparison was replicated across two different raters for a total of 1640 crowdsourcing tasks. Special care was given to make sure no single rater can perform more than 50 comparison tasks to simulate responses from a diverse population.

For the scaling experiment, we take the final ratings calculated using all 1640 task answers as a benchmark. To measure the change in accuracy as a function of the number of comparisons, we selected random subsets of the comparisons and fed them into the Elo system to arrive at ratings and predicted labels. Comparing these predicted labels to those calculated using the full suite of comparisons gives us an estimate of the accuracy. We also studied the effect of the number of items on the accuracy of classification by selecting samples of different sizes from the original 41 items, repeating the experiment and comparing the results.

Finally, while we do not conduct a large-scale survey to estimate the ground truth required for assessing the relative performance of the pairwise comparison and majority-voting methods, we provide confidence that our proposed approach yields reasonable results in the absolute by showing conversations and their estimated controversiality from our Twitter experiment. In Tab. 2, we present three conversations across the high, moderate, and low ranges of controversiality estimated using our proposed approach. These three conversations are specifically chosen from their respective ranges because they are succinct enough to be presented in tabular form, but all the the experiment data, including the Tweet IDs, the crowdsourcing responses, and their Elo ratings are available in a Git repository at https://github.com/hastinarimanzadeh/elo-paper-reproduction.

4.3.1 Ethical considerations. We have taken measures to ensure that we adhered to the necessary ethical considerations for running this crowdsourced experiment. While all tweets in the experimental data were public tweets, we anonymised the tweets such that no personally identifiable information from the posters was shown to the raters. When displaying each tweet, random usernames and random profile pictures were generated to replace the real ones and links and embedded images and videos in the tweets were redacted. Moreover, we did not gather any personal information from the raters. Finally, there was no experimenting or manipulation of information for the data presented to the raters and the entire process was minimally intrusive. The raters were paid 0.20$US per comparison for a target average hourly wage of 12.00$US. Based on our own pretests, we estimated each comparison to take approximately one minute to complete. Reviews from actual crowd raters on Turkerview ($n = 8$) show a median completion time of 44.5 seconds which translates to an average hourly wage of 16.18$US.

5 RESULTS

In Sec. 5.1 we discuss how the various facets of subjectivity quantified in Sec. 4.2 affect the accuracy of binary classification using the comparison and majority-vote methods, as expressed by their

---

[8]Tweets containing phrases "political climate" and "precip:" (standing for precipitation, used in reporting of daily weather forecast) as well as retweets were excluded.

Table 2. Climate twitter conversations across a range of controversiality. Replies to the top-level tweet are marked with arrows. Those highlighted in red ostensibly disagree with the top-level tweet. Higher rank means higher levels of controversiality based on Elo scoring of crowdsourced comparisons.

Rk.	Conversation
1	No, they're not. Every study in 2019 and prior concludes their ineffectiveness. Now that it's political they're suddenly effective. Science doesn't lie, but scientists with political agendas do. As we have already learned from climate 'scientists.'
	↪ The virus is spread through expectorant, meaning the stuff coming out of your nose and mouth when you sneeze and cough. That's what masks are meant to stop, and they're effective.
	↪ Maria, it's true, masks are effective.
	↪ How are you gonna do research on how effective are masks in 2019? It's 2020. This is covid 19. Something doctors have never studied before so everything you've supposedly "read" is not valid
20	Big News! Starbucks to ditch dairy for the environment. Starbucks admits that cow's milk is its biggest source of carbon dioxide emissions, yet it's still charging up to $0.80 more for lattes made with soy, coconut, or almond milk rather than dairy.
	↪ Hi I just have a query! I heard some nutritionist saying our body uses 80% of energy in digesting food is it true?
	↪ That is really great news! I would say the best news of 2020. Hope is it for all countries. It will serve as role model for others to change!
	↪ Almond milk is still pretty terrible for the environment...
	↪ If they're trying to promote environmental health, why charge more money for a latte? @Starbucks you can do better!
	↪ They have to bring in oat milk. Best for the environment and so good with coffee!!
40	It's snowing in New York City today. After two years in a tropical climate and three in a desert, it's a sight for sore eyes.
	↪ You weren't in Boise for Snowmageddon?
	↪ Just like being back in the mitten
	↪ Um, no...
	↪ I grew up in Jersey. I always loved the first snow. It's so beautiful. The lights shine on it and it looks like diamonds. Stay warm Paul.

f_1 scores. As we show below, we find that the comparison method yields higher f_1 scores than the 3-vote majority-vote method either when there is an equal number of comparisons and votes, or has the room to scale with increasing resource relative to majority-voting. We find that the comparison method is more robust to spammers and sources of unwanted discriminatory bias.

In Sec. 5.2, we proceed to simulate the scaling behaviour of the comparison method, computing and comparing its labelling accuracy as the number of items changes. We show that the number of required comparisons for a specific level of accuracy scales log-linearly with respect to the number of items to be labelled.

5.1 Simulation results

In order to compare the classification performance of the majority-vote method to the combination of comparison method and Elo system, we simulated a set of $N = 512$ items with true ratings

drawn from $\theta \sim \mathcal{N}(0, 1)$. Each item can be assigned a true binary label based on comparing their true ratings to the population mean $\theta > 0$. We also simulated a set of 100 raters, where each rater has a personal threshold drawn from $\theta_i^{\text{threshold}} \sim \mathcal{N}(0, 0.5)$, and their ambiguity in perception standard deviation and ambiguity in comparison are set to 0.5. The items are then classified with a majority-vote system with three votes by different raters, and subsequently with the comparison system presented in this paper, with $n_{\text{comparisons}}$ comparisons randomly selected sample from the pool of all possible comparisons. Each method then predicts a binary label for each item. The accuracy of the two methods is calculated in the form of the positive label f_1 score, where for every item the label predicted by each method is compared to its true label.

At each step, one of the parameters of the simulation (from the first five parameters in Tab. 1) is modified and the results are studied. The simulations for each parameter are repeated multiple times, and we compute the mean and the 5σ confidence intervals of the values of interest (often the difference in f_1 scores of the two methods, $f_1^{\text{comparison}} - f_1^{\text{majority}}$). These results are summarised in the figures that follow.

For most of these figures, we plot the difference in f_1 scores either on the y-axis or the z-axis of a surface. Regions above zero indicate better performance by the comparison method. We have plotted a vertical line at the x-axis of 1, which is when there is an equal number of votes to comparisons. This location is a useful indicator, as it is generally speaking when an equal amount of resources are given to the two approaches, but should not be taken as the only fair comparison point. Instead, we understand it to be the comparison method's worst relative performance, as the majority-vote method tends to be more limited in terms of scaling with additional resources, which we show in Fig. 6 (right panel). On the other hand, the number of possible comparisons exceeds the 3-vote majority-vote for $N > 8$, moving us to the region of the figures to the right of $x = 1$.

5.1.1 Effects of Subjectivity. In this section, we analyse the relative performance of the majority-vote method and the comparison method by measuring the difference in binary labelling performance, as indicated by f_1 scores. A positive value of difference in f_1 score indicates better performance of the comparison method in binary classification compared to the majority-vote method. To this end, we simulate 3-vote majority-voting and comparison tasks while varying values for rater threshold diversity, ambiguity in comparison, and ambiguity in perception. Sections 5.1.2 and 5.1.3 will extend these simulations to $2n + 1$-vote majority-voting and rater populations with implicit biases.

Diversity in threshold. We begin by looking at the first type of translational subjectivity, which is the diversity in raters' thresholds for identifying an item to be possessing the construct in question (e.g. how much toxicity is toxic?) which is parameterised by the level of variability σ in the threshold values in the rater population. As previously discussed, this type of translational subjectivity factors into item-wise rating tasks, and should therefore favour the comparison method over the majority-vote method. Our simulation results, presented in Fig. 4, confirm that higher degrees of subjectivity in the form of greater rater threshold diversity confers an advantage to the comparison method. The heat map shows the difference in the two methods' f_1 scores, with the blue regions indicating better performance by the comparison method. These findings are also presented in Fig. 5(c), which shows f_1 score differences at specific values of σ.

Ambiguity in comparison. Next, we consider the other type of translational subjectivity, ambiguity in comparison, which factors only into comparison tasks. It stands to reason that it should negatively impact the performance of the comparison method, therefore favouring the majority-vote method. This is indeed generally the case, as shown in Fig. 5(b). However, interestingly, we also see that a moderate amount of ambiguity in comparison actually improves the accuracy of the comparison
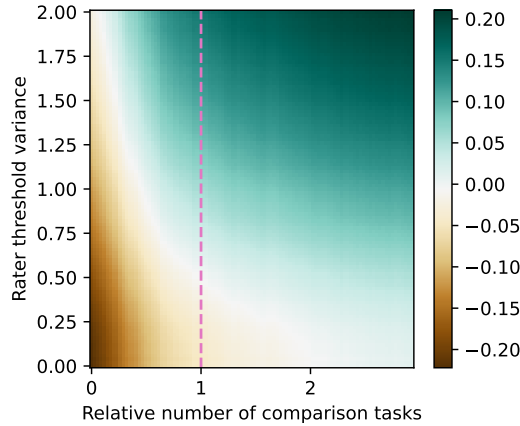
Fig. 4. Subjectivity of a problem, expressed in terms of diversity of threshold among raters, has a positive effect on the relative efficacy of the pairwise comparison method. The heat map shows the difference in positive label f_1 scores of the pairwise comparison method and the majority-vote method. The blue regions on the heat map (i.e. positive-valued) display areas where the comparison method performs better than the majority-vote method. The dashed vertical line indicates the situation where equal numbers of comparison tasks and majority-vote tasks are performed.

method. Specifically, the performance of the comparison method improves as the ambiguity in comparison is increased to 0.8, after which its performance drops back down. This is due to the fact that in our implementation of the Elo rating system, when the raters are unable to differentiate between two items, they are allowed to declare them as equal, which results in the ratings of the two items converging toward each other. Past a certain point, however, large values of ambiguity in comparison mean even items with a large disparity in true ratings cannot be distinguished from each other, thereby resulting in less information made available to the rating system. For example, when the ambiguity in comparison has a value of 2, an item in the third percentile in terms of true rating is deemed indistinguishable from an item in the 97th percentile.

Ambiguity in perception. Turning to definitional subjectivity, Fig. 5(a) illustrates how the comparison and the majority-vote methods perform relative to each other for six different values of ambiguity in perception. We see that when there is an equal number of comparisons to votes, the two approaches perform the same. However, as discussed in Sec. 5.1, the majority-vote method is capped at $3N$ tasks unless a larger number of votes are cast per item, which does not scale well. The region for our assessment, then, is to the right of the vertical pink line, where larger values of perception ambiguity result in the comparison method having a larger lead in accuracy to that of the majority-vote method. This lead is accentuated as the number of comparison tasks performed increases.

5.1.2 Rater bias. In this setup, each item has an additional binary observable representing a *discriminatory feature* (e.g., racial features in the profile photo) which ideally should not be affecting the assessment of an item. The raters may, nonetheless, take it into consideration in their assessment. If the weight distribution associated with the binary feature has a mean of zero in the rater population, this will simply induce a larger error in estimated ratings for the items exhibiting that feature. On the other hand, bias is introduced to the classification if the raters are more likely to interpret the existence of this additional feature in one direction. This happens when the weight distribution associated with the discriminatory feature has a nonzero mean, e.g., on average seeing
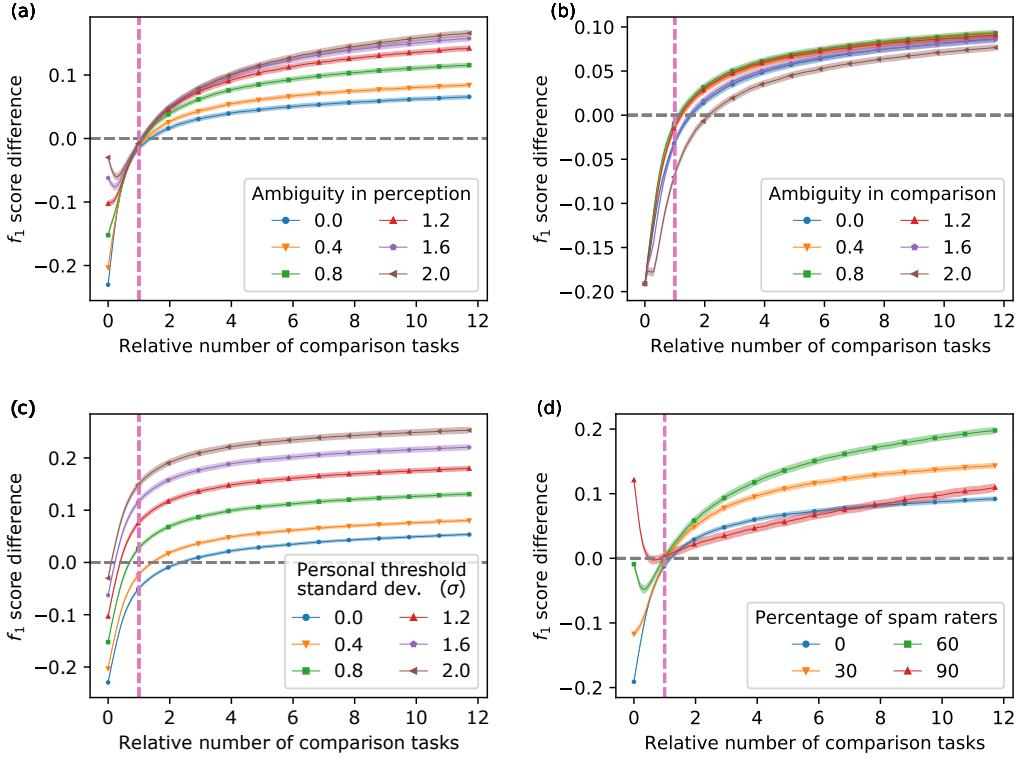
Fig. 5. f_1 score difference ($f_1^{\text{comparison}} - f_1^{\text{majority}}$) of the comparison and majority-vote methods for parameters ambiguity in perception (a), ambiguity in comparison (b), personal threshold differences between raters (c) and percentage of spam raters (d) as the relative number of comparison tasks, that is, the number of comparison tasks divided by the number of the majority-vote tasks, increases. A positive f_1 score difference (shown on the y-axis) indicates an advantage in using the comparison method over the majority-vote method in terms of binary labelling accuracy with a given ratio of comparison tasks to majority-vote tasks. The x-axis indicates the ratio of comparison tasks to majority-vote tasks, i.e. a value of 2 indicates performing twice as many comparisons as votes, where the number of majority-vote tasks is fixed to 3 tasks (votes) per item. The vertical line at x=1 indicates performing the same number of comparison tasks and majority-vote tasks. Larger values of ambiguity in perception and higher variance of personal threshold, as with medium amounts of ambiguity in comparison and spammer raters, result in a higher relative efficacy of the comparison method compared to the majority-voting method.

a black profile picture makes raters more likely to assess the comment as toxic regardless of the content.

To simulate this, we add to each item an additional binary $0, 1$ feature with probability 0.5. The weight associated with this feature is distributed in population as a skewed beta distribution with parameters $\alpha = 2$ and $\beta = 16$ translated and scaled to have support of $[-1, +1]$, i.e. $\omega_d \sim 2Beta(2, 16) - 1$. This implies that the raters are on average more likely to associate the exhibition of the discriminatory feature to lower ratings and therefore the negative label. The results in Fig. 6 (left panel) show that while the comparison method still reflects the bias against the discriminatory

feature that exists among the raters, the magnitude of the exhibited bias is generally quite smaller than in the majority-vote method.
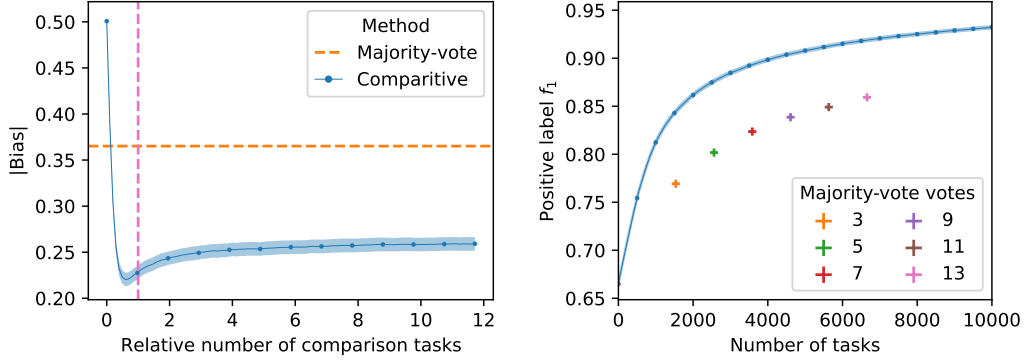

Fig. 6. Left: The magnitude of bias in the binary classification of items with an observable discriminatory feature for the comparison and majority-vote methods with the same set of raters. The bias is calculated as the mean difference between the predicted label (using either the comparison or majority-vote method) and the true labels of all items that exhibit the biased feature. With the same group of raters, items exhibiting the discriminated feature are less likely to be classified with the negative label using the comparison method as compared to the majority-vote method. Right: The efficacy of the comparison method for a subjective task in a diverse population of simulated raters with personal threshold variance of 1.0, as compared to majority-voting method with different number of votes per item. The comparison method consistently out-performs the majority-vote method. As the number of votes of majority voting increases, the difference becomes slightly less substantial compared to the commonplace 3-vote majority voting.

5.1.3 Additional results. In addition to assessing the impact of subjectivity on the two methods, we conducted additional analysis on how the methods perform when faced with spam, and when the number of votes increases in the majority-vote method.

The effect of spam. For this analysis, we simulated the effect of spammers by having a certain percentage of raters give random responses to the comparisons or absolute vote tasks. Our results are presented in Fig. 5(d). Our findings here are similar to the results for ambiguity in perception, where while the two methods perform the same when there is resource parity, the comparison method can scale, making the relevant region for our assessment to the right of the vertical pink line. Here, we see that the comparison method is more robust to spamming than the majority-vote.

Increasing voters per majority-vote task. It is possible to run the majority-vote method with more than 3 votes. Fig. 6 (right panel) compares the expected f_1 scores of the majority-vote method with different numbers of votes per item, to those of the comparison method with the same number of tasks.

Through the simulations analysed in this section, we are able to see how the comparison method outperforms the majority-vote in the increasing presence of subjectivity, namely with diversity in threshold and ambiguity in perception. Further, we find that while higher ambiguity in comparison has an adverse effect on the performance of the comparison method, a moderate amount in fact proves to be beneficial. We subsequently showed that the comparison method is more robust to rater bias and spam when compared to the majority-vote method.

There is nevertheless a trade-off between the accuracy of the comparison method and the number of pairwise comparisons. We study this trade-off and how the number of required comparisons and labelling accuracy scale with the total number of items in the following section.

5.2 Scaling behaviour

To understand the scaling behaviour of the comparison method, we can compare the labelling accuracy of the comparison method both in simulation and in empirical observation with different numbers of items. Take for example a set of N items. For different numbers of comparisons ($n_{\text{comparisons}}$) we plot the average positive label f_1 score. If we repeat the same for different values of N, we arrive at a family of trajectories. These trajectories collapse on top of each other after applying a scaling function of N to the number of comparisons (the horizontal axis) for each curve. This scaling function can tell us how one needs to increase the number of comparisons when increasing the system size N for achieving the same level of quality.

The same analysis was repeated on the experimental data described in Sec. 4.3, with the key difference being that true ratings and labels are not available for the experimental items. Instead, we used the Elo rating based on the full set of all 1640 performed comparisons as the comparison baseline. Different system sizes can be simulated by selecting a subset of items with size N, and removing all comparisons that include items not in this subset. For each system size, positive and negative "true" labels were assigned based on comparing the full-set rating of each item in the subset with the subset median rating. For each system size, many random samples of comparisons with different size $n_{\text{comparisons}}$ were selected, and classification labels were predicted based on Elo ratings calculated from these comparisons. For each system size, this process is repeated many times. Figure 7(c) shows the change in f_1 score as a function of $n_{\text{comparisons}}$ and N, while Fig. 7(d) shows that the same trajectories collapse after correcting for our hypothesised $O(N \log N)$ scaling, confirming the hypothesis that the number of required comparisons for the same level of labelling quality grows log-linearly with respect to the number of items.

These results, as well as those based on the mathematical modelling presented in Sec. 5.1, lay a road-map for using our proposed comparison method for crowdsourcing tasks intended to measure subjective constructs. In the next section, we summarise our results and present our recommendations for real-world implementation of our proposed method.

6 DISCUSSION AND CONCLUSION

In this paper, we proposed a comparison-based labelling approach for subjective constructs which combines pairwise comparisons with Elo scoring. We showed through simulations that the comparison method fares better against the ubiquitous majority-vote method in dealing with most types of subjectivity, measured as reductions in misclassification error and classification bias.

Specifically, using an agent-based model of crowdsourced labelling to quantify different forms of task subjectivity, we compared the performance of the two labelling approaches. We found that the comparison method either is better at reducing random error (i.e. achieving higher f_1 scores) when the number of comparisons and votes are equal (i.e. under resource parity), or has more room to improve with an increasing number of tasks whereas the majority-vote tends to become capped. Second, we incorporated a discriminatory feature into our agent-based model to test how the two approaches performed in bias reduction. We find a nonzero value of bias in both methods, which demonstrates that bias reduction needs to be accomplished via other steps in the human-labelling pipeline. However, the comparison method outperforms the majority-vote in that it does not inflate bias in the way majority-voting does.

Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error
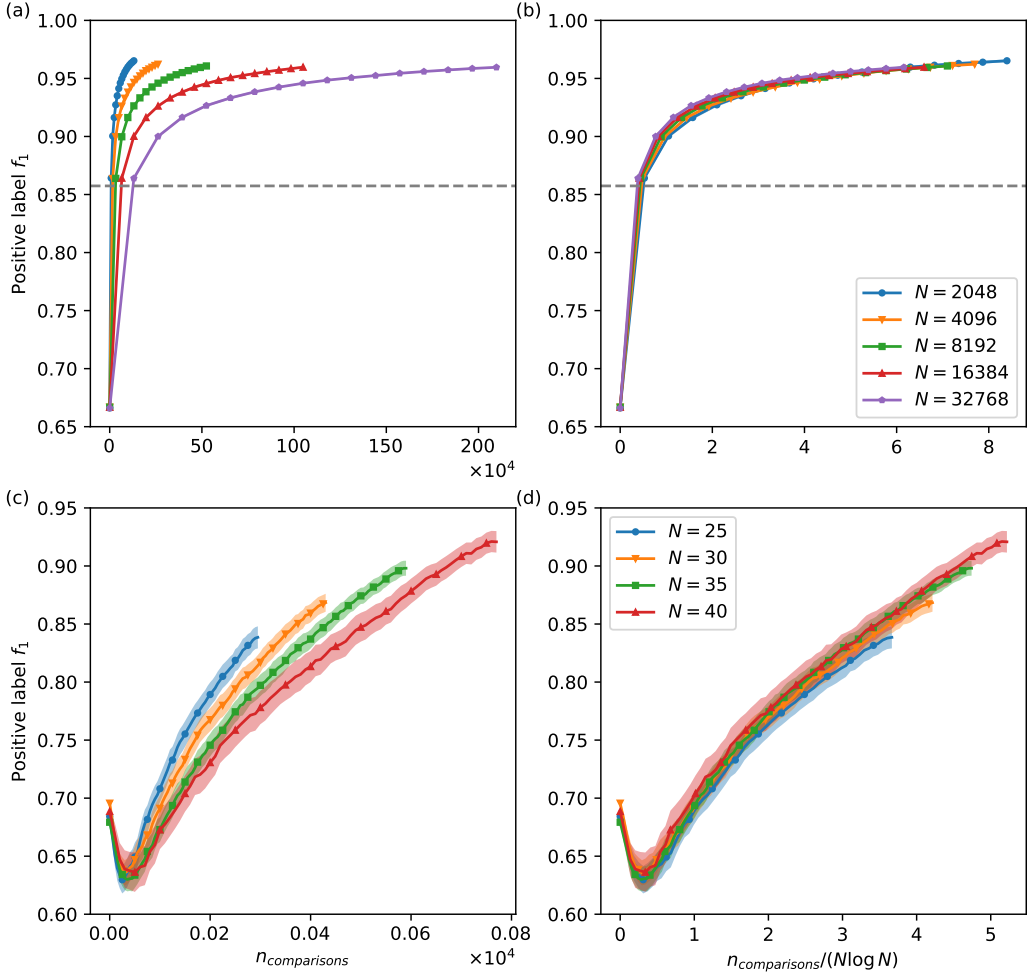Compared to the Majority-vote Method

334:21

Fig. 7. The number of required comparisons for achieving desired level of accuracy, as denoted by the positive label f_1 score, in a subjective binary classification task scales log-linearly $O(N \log N)$ with the number of items being classified with the comparison method N. This can be verified by the scaling analysis on (b) simulations described in Sec. 4.2 and (d) empirical analysis on real-world subjective task described in Sec. 4.3. The collapse of the trajectories, after correcting for the hypothesised scaling function $N \log N$ on the horizontal axis, validates the hypothesised scaling function. The hyphenated horizontal lines in (a) and (b) denote the expected positive label f_1 score for the 3-vote majority-voting classification method. The shaded area indicates the 5-σ confidence interval where σ is the standard error of mean.

6.1 Considerations when using the comparison method in real-world studies

To facilitate applied work, we conducted a set of comparison-based crowdsourcing tasks with Twitter conversations to study how the number of comparison-based tasks scales with respect to the number of to-be-labelled items. Based on this experiment, we arrived at an $O(N \log N)$ scaling of the required random comparisons. Our results show that the previously assumed relationship

between the number of comparison tasks and items in the literature (i.e. $O(N^2)$) overestimated the resource intensity of the comparison method [4]. Instead, our observed log-linear scaling means that comparison-based labelling is more tractable in large-scale studies than previously thought.[9]

In Appendix A, we provide a step-by-step guide for implementing our proposed pipeline and also introduce our open-source reference implementation of the Elo rating system. Here, we briefly discuss some additional considerations that concern crowdsourced labelling tasks in general.

First, regardless of which labelling method, it is important to keep in mind the representativeness of the crowd worker sample to the desired population. Following our discussion in Sec. 2, the target measure when working with subjective constructs is the mean population perception μ_θ of a given item, which is sensitive to sampling bias. We did not address the question of sampling bias in our study, but point to the survey sampling literature that provides instruction on how to address these concerns [20, 21].

Second, subjectivity-induced variation aside, crowdsourced labels can vary in quality so quality control is an important consideration. In our study we explored the effect of non-malicious spam and otherwise random labels, which enters into our theoretical framework in the same manner as ambiguity in perception. We show that spam mitigation for the comparison method scales better with resources relative to the majority-vote. We also discuss spam detection methods in Appendix A.1. We do not address how to account for crowd raters who tactically attempt to sway the results, which we believe should be minimal for innocuous tasks where there is no internal incentive for maliciousness. However, we recognise it is a potentially important consideration when working with politically sensitive topics.

Related to this final point, we note that our proposed pipeline is modular, which makes it easy for components to be customised for the specific needs of a study. For example, subsystems designed to detect and mitigate intentionally malicious workers can be plugged into the pipeline between the labelling and aggregation tasks in a straightforward manner.

6.2 Limitations and future work

We conclude by discussing avenues of future work that address limitations of our current study. One main area of future exploration stem from our synthetic design and largely modular testing framework. First, we have tested different types of subjectivity independently of each other instead of also considering how they may interact. Measuring and testing potential interactions between various aspects of subjectivity we delineate in our model is a difficult task and will requiring novel study design and a broad data-gathering effort. Part of this requires future work to move toward real-world data, which lets us assess the holistic performance of different crowdsource labelling methods, as it is difficult to directly compare the magnitude of different types of subjectivity in a synthetic setting.

Relatedly, by focusing on only the labelling and aggregation components, we have largely assumed away how other important considerations in a crowdsource labelling pipeline (e.g. crowd-worker representativeness, bias reduction, and spam and gaming mitigation) can interact with the labelling and aggregation steps. For example, in terms of bias reduction, although we show that the comparison method is less susceptible to discriminatory features compared to the majority-vote method, we do not propose a solution to correct it. An active measure might be required on part of the researcher to eliminate or account for the biases introduced by the possibly discriminatory

---

[9]This allows researchers to run small-scale pilots until the desired accuracy of results are reached, after which comparison tasks can be scaled up to an arbitrary number of items, with the number of required comparisons growing as $O(N \log N)$ for the same accuracy level.

Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error
Compared to the Majority-vote Method

334:23

features of the items. This is an active area of research [37, 60], which we believe can be fruitfully combined with work such as ours that focus on labelling and aggregation in crowdsourcing tasks.

Some additional extensions of our work are worth exploring. First, while we tried to include many aspects of subjectivity in our simple mathematical model of human understanding, there may exist other facets to subjectivity that we have not incorporated into our formalisation. For example, in this study, we assumed unimodal distributions for weights, which recent work has expanded beyond [3]. Second, we do not provide analytical proof of the scaling relationship $O(N \log N)$. Rather, we took an empirical approach to arrive at a scaling relationship. We believe that an analytical derivation of this scaling relationship would prove a valuable addition to this discussion. Finally, while we worked with binary measures in this study, the comparison method and the Elo rating system directly allow for ordinal or even interval measures. These can be used instead of the binary label to train a regression model or to study the correlation of item ratings with time or across groups. As such, our work likely understates the performance of the comparison method relative to item-wise labelling which requires extended task designs to allow for nonbinary labels, and are likely prone to higher levels of translational subjectivity. Much more work can be done in this area to understand these potential outputs and how they perform in capturing the target subjective construct. We hope our present work provides a useful road-map for future work.

ACKNOWLEDGMENTS

REFERENCES

[1] Omar Alonso, Catherine C Marshall, and Marc Najork. 2015. Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. 101–110. https://doi.org/10.1145/2756406.2757741

[2] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 1100–1105. https://doi.org/10.1145/3308560.3317083

[3] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

[4] James P Bagrow. 2020. Democratizing AI: non-expert design of prediction tasks. *PeerJ Computer Science* 6 (2020), e296. https://doi.org/10.7717/peerj-cs.296

[5] Agathe Balayn and Alessandro Bozzon. 2019. Designing evaluations of machine learning models for subjective inference: the case of sentence toxicity. *arXiv preprint* (2019). https://doi.org/10.48550/arXiv.1911.02471

[6] Ali Borji. 2019. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding* 179 (2019), 41–65.

[7] Guillermo F Cabrera, Christopher J Miller, and Jeff Schneider. 2014. Systematic labeling bias: De-biasing where everyone is wrong. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 4417–4422.

[8] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or There. In *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings (Lecture Notes in Computer Science, Vol. 4956)*. Springer, 16–27. https://doi.org/10.1007/978-3-540-78646-7_5

[9] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346. https://doi.org/10.1145/3025453.3026044

[10] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202. https://doi.org/10.1145/2433396.2433420

[11] Justin Cheng and Michael S Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 600–611. https://doi.org/10.1145/2675133.2675214

[12] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25. https://doi.org/10.1145/3359164

[13] Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. *Advances in neural information processing systems* 28 (2015), 1486–1494.

[14] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch (CEUR Workshop Proceedings, Vol. 842)*. CEUR-WS.org, 26–30.

[15] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 67–73. https://doi.org/10.1145/3278721.3278729

[16] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 32–41. https://doi.org/10.1609/hcomp.v4i1.13270

[17] Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *European Semantic Web Conference*. Springer, 701–710. https://doi.org/10.1007/978-3-319-18818-8_43

[18] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. Crowdsourcing ground truth for medical relation extraction. *arXiv preprint arXiv:1701.02185* (2017). https://doi.org/10.1145/3152889

[19] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[20] Yuxiang Gao, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2021. Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis* 16, 3 (2021), 719–744.

[21] Andrew Gelman and Thomas C Little. 1997. Poststratification into many categories using hierarchical logistic regression. (1997).

[22] Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist* 52, 1 (1997), 45. https://doi.org/10.1037/0003-066X.52.1.45

[23] Richard D Goffin and James M Olson. 2011. Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science* 6, 1 (2011), 48–60.

[24] William T Goldberg. 2010. Two nations, one web: Comparative legal approaches to pornographic obscenity by the United States and the United Kingdom. *BUL Rev.* 90 (2010), 2121.

[25] Yuan Guo, Peng Tian, Jayashree Kalpathy-Cramer, Susan Ostmo, J. Peter Campbell, Michael F. Chiang, Deniz Erdogmus, Jennifer G. Dy, and Stratis Ioannidis. 2018. Experimental Design under the Bradley-Terry Model. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. ijcai.org, 2198–2204. https://doi.org/10.24963/ijcai.2018/304

[26] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Miguel Bordallo López, Vassilis Kostakos, and Jorge Goncalves. 2020. Augmenting Automated Kinship Verification with Targeted Human Input. In *PACIS*. 141.

[27] Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 377–382.

[28] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3290605.3300637

[29] Gerardo Iñiguez, Carlos Pineda, Carlos Gershenson, and Albert-László Barabási. 2022. Dynamics of ranking. *Nature Communications* 13, 1 (2022), 1–7. https://doi.org/10.1038/s41467-022-29256-x

[30] Ikbeom Jang, Garrison Danley, Ken Chang, and Jayashree Kalpathy-Cramer. 2022. Decreasing Annotation Burden of Pairwise Comparisons with Human-in-the-Loop Sorting: Application in Medical Image Artifact Rating. *arXiv preprint* (2022). https://doi.org/10.48550/arXiv.2202.04823

[31] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648. https://doi.org/10.1145/2818048.2820016

[32] Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. *arXiv preprint arXiv:2004.14667* (2020).

[33] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative*

work. ACM, 1301–1318. https://doi.org/10.1145/2441776.2441923

[34] James Q Knowlton. 1966. On the definition of "picture". *AV communication review* 14, 2 (1966), 157–183. https://doi.org/10.1007/BF02769550

[35] Adriana Kovashka and Kristen Grauman. 2013. Attribute adaptation for personalized image search. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society, 3432–3439. https://doi.org/10.1109/ICCV.2013.426

[36] Li Kuang, Huan Zhang, Ruyi Shi, Zhifang Liao, and Xiaoxian Yang. 2020. A spam worker detection approach based on heterogeneous network embedding in crowdsourcing platforms. *Computer Networks* 183 (2020), 107587.

[37] Yanying Li, Haipei Sun, and Wendy Hui Wang. 2020. Towards fair truth discovery from biased crowdsourced answers. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 599–607. https://doi.org/10.1145/3394486.3403102

[38] Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems* 31 (2018).

[39] OO Marchenko, OS Radyvonenko, TS Ignatova, PV Titarchuk, and DV Zhelezniakov. 2020. Improving text generation through introducing coherence metrics. *Cybernetics and Systems Analysis* 56, 1 (2020), 13–21.

[40] Joshua E Menke and Tony R Martinez. 2008. A Bradley–Terry artificial neural network model for individual ratings in group competitions. *Neural computing and Applications* 17, 2 (2008), 175–186.

[41] Vikram Mohanty, Kareem Abdol-Hamid, Courtney Ebersohl, and Kurt Luther. 2019. Second opinion: Supporting last-mile person identification with crowdsourcing and face recognition. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 86–96. https://doi.org/10.1609/hcomp.v7i1.5272

[42] Vikram Mohanty, David Thames, and Kurt Luther. 2018. Are 1,000 features worth a picture? combining crowdsourcing and face recognition to identify civil war soldiers. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*. https://par.nsf.gov/biblio/10081893

[43] Charles Kay Ogden and Ivor Armstrong Richards. 1925. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Vol. 29. Harcourt, Brace.

[44] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *2011 International Conference on Computer Vision*. IEEE, IEEE Computer Society, 503–510. https://doi.org/10.1109/ICCV.2011.6126281

[45] Nathaniel D Porter, Ashton M Verdery, and S Michael Gaddis. 2020. Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PloS one* 15, 6 (2020), e0233154.

[46] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. 2011. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*. IEEE, IEEE, 3097–3100. https://doi.org/10.1109/ICIP.2011.6116320

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[48] Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 213–217.

[49] Joni Salminen, Ahmed Mohamed Kamel, Soon-Gyo Jung, and Bernard Jansen. 2021. The Problem of Majority Voting in Crowdsourcing with Binary Classes. In *Proceedings of 19th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

[50] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gyo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 88–94. https://doi.org/10.1109/SNAMS.2018.8554954

[51] Joni O Salminen, Hind A Al-Merekhi, Partha Dey, and Bernard J Jansen. 2018. Inter-rater agreement for social computing studies. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 80–87.

[52] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

[53] Sarah Sobieraj and Jeffrey M Berry. 2011. From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication* 28, 1 (2011), 19–41.

[54] Hal Stern. 1992. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences* 23, 1 (1992), 103–117. https://doi.org/10.1016/0165-4896(92)90040-C

[55] Jane Lawrence Sumner, Emily M Farris, and Mirya R Holman. 2020. Crowdsourcing reliable local data. *Political Analysis* 28, 2 (2020), 244–262.

[56] Takeru Sunahase, Yukino Baba, and Hisashi Kashima. 2017. Pairwise hits: Quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[57] Fangna Tao, Liangxiao Jiang, and Chaoqun Li. 2020. Label similarity-based weighted soft majority voting and pairing for crowdsourcing. *Knowledge and Information Systems* 62, 7 (2020), 2521–2538.

[58] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research* 72 (2021), 1385–1470. https://doi.org/10.1613/jair.1.12752

[59] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*. ACM, 21–26.

[60] Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A Leiva. 2022. Debiased Label Aggregation for Subjective Crowdsourcing Tasks. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8. https://doi.org/10.1145/3491101.3519614

[61] Shaun Wallace, Brendan Le, Luis A Leiva, Aman Haq, Ari Kintisch, Gabrielle Bufrem, Linda Chang, and Jeff Huang. 2020. Sketchy: Drawing inspiration from the crowd. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27. https://doi.org/10.1145/3415243

[62] Jiele Wu, Chau-Wai Wong, Xinyan Zhao, and Xianpeng Liu. 2021. Toward effective automated content analysis via crowdsourcing. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, IEEE, 1–6. https://doi.org/10.1109/ICME51207.2021.9428220

[63] Peng Ye and David Doermann. 2013. Combining preference and absolute judgements in a crowd-sourced setting. In *ICML Workshop*. 1–7.

[64] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552. https://doi.org/10.14778/3055540.3055547

[65] James Zou, Kamalika Chaudhuri, and Adam Kalai. 2015. Crowdsourcing feature discovery via adaptively chosen comparisons. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 3. 198–205. https://doi.org/10.1609/hcomp.v3i1.13231

A APPENDIX: USING THE COMPARISON METHOD IN THE REAL-WORLD

In this paper, we discussed merits and properties of the comparison method for labelling subjective constructs. In this section we will provide a brief summary of our recommended approach as a step-by-step guide, as well as briefly introduce methods of dealing with spamming or inattentive raters.

Let us assume that you have a set of items and an intended subjective construct in mind. This approach is designed for subjective qualities where one can say A has more (or less) of this quality than B, such as the toxicity of tweets or realisticness of an AI generated image.

Depending on the question at hand, you need a representative sample of items to be compared. If you intend to compare relative toxicity of all English language tweets on the subject of climate change compared to a baseline month, you need a representative sample of English climate change tweets. Similarly, attention has to be paid on selecting the rater population, as the rating and labels derived from this method is representative of the opinions and the viewpoint of the raters.

The next step is to run an exhaustive pilot, similar to Fig. 7(d), for a smaller subset of your sample. For example, for gathering a larger dataset of controversial and non-controversial tweets for training machine-learning models or to study the evolution of controversiality of climate tweets across time, researchers might consider running a pilot experiment similar to what was described in Sec. 4.3. Select a sub-sample of your items and perform all possible comparisons between pairs of items. At this step you can elect to perform each possible comparison multiple times by different raters, this also allows you to compare their answers using existing rater agreement statistics.

Plot the result of this pilot study in the same manner as Fig. 7(d). The scaling trajectory of the pilot experiment would allow you to select the appropriate compromise between labelling accuracy and number of items labelled, based on the budget allocated for the study. Let us say you require that the items in your final output have a labelling accuracy, as indicated by f_1 score, of at least 0.9.

Crowdsourcing Subjective Annotations Using Pairwise Comparisons Reduces Bias and Error
Compared to the Majority-vote Method

334:27

The x-axis value where the scaling trajectory intersects $y = 0.9$ predicts a value for $\frac{n_{\text{comparisons}}}{N \log N}$ where N stands for the number of items and $n_{\text{comparisons}}$ for the number of comparisons. If your budget allows for a certain maximum number of crowdsourcing tasks, solving for N gives you the number of items that you can label. For the experiment performed in Sec. 4.3, to get a f_1 score of 0.8 for 1000 unique conversations for this specific task, you can expect to need close to 16000 comparisons based on the scaling plot Fig. 7(d).

Note that the trajectory of the scaling plot is subject to the properties of the tasks at hand, meaning that a different research question or a significantly different population of raters might require its own specific pilot study and that task subjectivity parameters such as ambiguity in comparison and perception can affect the number of required pairwise comparisons and the result of the Elo system for a specific value of accuracy. As we do not propose a direct method for measuring these properties, the most reliable approach of ensuring the efficacy of the full-scale experiment is to determine the number of required comparisons for N items to achieve high-quality results by conducting pilot experiments.

After determining the number of items N in the full-scale experiment based on the budget and the pilot experiment results, select a sample of size N from the items and perform random comparisons between pairs of items. Aggregating the results through the Elo rating system gives you a rating for each item, indicating the strength of subjective construct in that item as construed by the average rater. This rating can be translated into binary classification labels.

For both the pilot and the main study, we recommend cycling through all comparisons in random order for multiple epochs, until Elo ratings of each item arrives at a stable value. This can be viewed similar to the practice of cycling through the training data multiple times when training a deep neural network model, which here helps ease the negative effects of selecting a small value for parameter k of the Elo method.

If your research question depends on comparing one population of items to another, let us say comparing toxicity of English climate change tweets to background level of toxicity of English in each month, you can draw separate samples for each population and calculating ratings separately for the baseline group and benchmark group. Subsequently, you can "anchor" the ratings from the baseline group to the benchmark group by, e.g., selecting the 5 items closest to the average in each month in the baseline group, performing comparisons using a binary-search algorithm for each of the 5 baseline tweets against sorted benchmark tweets belonging to the same month. This would give you an approximate ranking (and rating) for each of 5 baseline tweets among the benchmark tweets of the same month that can be used for labelling items in the benchmark set.

A.1 Detecting spammers

While, as we showed in Sec. 5.1.3, the comparison method overall is more robust toward inattentive raters or spammers compared to the majority-vote method, it might be necessary to have the ability to detect flagrant instances of spamming. While research in this field is still in its preliminary stage, we suggest multiple mitigation approaches. The most simple approach relies on the correlation between the difference in rating $R_A - R_B$ and the selected outcome S_A in every comparison performed by one worker. We expect an attentive rater who has performed a sufficiently large number of comparisons to show a significant positive correlation, and the absence thereof or a negative correlation can be interpreted as a sign of possible spamming. Another approach is based on the probability of the selected outcome for each comparison performed by one rater. Equation. (2) can be used to calculate a probability for the selected outcome of a comparison by a rater. Consistently selecting low-probability outcomes, as can be indicated by median selection probability, can signify possible spamming.

In all of the above-mentioned methods, when implementing these approaches to investigate a specific rater, it is advisable to calculate ratings and rankings based on the contributions of all raters except the rater being investigated.

Reliability of these approaches can be improved by having a larger set of comparisons performed by each rater, though this comes at a cost of reducing diversity among the raters. It is also possible to perform attention-check comparisons, consisting of items known to have large rating differences based on the pilot results. The results of these comparisons, when combined with each rater's comparisons can greatly bolster the previously mentioned spammer-detection approaches. Implementation of such an approach can be akin to the one from Hettiachchi et al. [26], where their crowdsourced kinship verification had a number of categorically verifiable kinship queries, specifically in the form of asking whether a cartoon character looks like or is related to a human.

A.2 Elo rating system reference implementation

Our implementation of the Elo rating system is available as a Python package. This package can be installed using the command `python -m pip install -U elo-rating`. The documentation as well as the source code can be found at https://github.com/hastinarimanzadeh/elo-rating. Listing 1 displays common usage pattern as a simple code snippet.

Listing 1. Example use-case of the Python Elo rating system implementation in aggregating comparisons.

```python
from elo_rating import Elo

# create an elo class, denom referes to 1/(logistic growth rate),
# i.e. the number 400 in Eq. 2
e = Elo(denom=400)

# add comparisons one by one
e.add_match("p1", "p2", 1.0, k=30, default_rating=1400)

# or many at the same time
e.add_matches([("p1", "p2", 1.0), ("p2", "p1", 0.5)],
  k=30, default_rating=1400)

# get a list of all seen items
e.items() # ['p1', 'p2']

# item ratings based on current comparisons
e.ratings()
# => {'p1': 1428.3358360702414, 'p2': 1371.6641639297586}

# item ratings based on current comparisons
e.rankings()
# => {'p1': 0, 'p2': 1}

e.ranking('p1') # 0
e.rating('p1') # 1428.3358360702414

# probability of win for p1 in a hypothetical match with p2
e.expected_score("p1", "p2") # 0.5353606653429002
```