frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 01 April 2022
doi: 10.3389/frai.2022.818451

Scaling and Disagreements: Bias, Noise, and Ambiguity

Alexandra Uma[1]*, *Dina Almanea*[1] and *Massimo Poesio*[1,2,3]

[1] Computational Linguistics Lab, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom, [2] Digital Environment Research Institute, Queen Mary University of London, London, United Kingdom, [3] Turing Institute, London, United Kingdom

Crowdsourced data are often rife with disagreement, either because of genuine item ambiguity, overlapping labels, subjectivity, or annotator error. Hence, a variety of methods have been developed for learning from data containing disagreement. One of the observations emerging from this work is that different methods appear to work best depending on characteristics of the dataset such as the level of noise. In this paper, we investigate the use of an approach developed to estimate noise, temperature scaling, in learning from data containing disagreements. We find that temperature scaling works with data in which the disagreements are the result of label overlap, but not with data in which the disagreements are due to annotator bias, as in, e.g., subjective tasks such as labeling an item as offensive or not. We also find that disagreements due to ambiguity do not fit perfectly either category.

Keywords: overlapping labels, annotation disagreement, observer disagreement, temperature scaling, model calibration, cost-sensitive loss

OPEN ACCESS

Edited by:
Matt Lease,
University of Texas at Austin,
United States

Reviewed by:
Christopher Welty,
Google, United States
Alexander Braylan,
University of Texas at Austin,
United States

***Correspondence:**
Alexandra Uma
alexandra.uma2@gmail.com

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *19 November 2021*
Accepted: *01 March 2022*
Published: *01 April 2022*

Citation:
Uma A, Almanea D and Poesio M (2022) Scaling and Disagreements: Bias, Noise, and Ambiguity. Front. Artif. Intell. 5:818451. doi: 10.3389/frai.2022.818451

1. INTRODUCTION

Crowdsourced data are often rife with disagreements between coders. Hence, a variety of methods have been developed for learning from data containing disagreement. In a previous study, the focus was on developing methods for removing items on which annotators disagreed (Beigman-Klebanov and Beigman, 2009), or aggregation methods able to learn "ground truth" from such data (Dawid and Skene, 1979; Smyth et al., 1994; Carpenter, 2008; Whitehill et al., 2009; Hovy et al., 2013) (see, e.g., Sheshadri and Lease, 2013; Paun et al., 2018, 2022; Uma et al., 2021b for review). More recent work however suggests that better results are obtained by methods training directly from data containing disagreements (Raykar et al., 2010; Rodrigues and Pereira, 2017; Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021; Uma et al., 2021b). But another finding emerging from this recent work is that different methods for learning from data containing disagreements work best depending on the dataset (Uma et al., 2021b). One possible explanation for this difference in performance is disagreements can be due to a number of causes, ranging from annotator error to problematic annotation schemes (e.g., with overlapping labels) to genuine item ambiguity to more general item difficulty. An early proposal regarding distinguishing between different types of disagreement was made by Reidsma and Carletta (2008), who showed that disagreements due to **(random) noise**—random annotator errors—affect model training differently from disagreements due to **bias**—annotator-dependent patterns. Such work raises the question of whether it is possible to distinguish between these two types of disagreement (or other types perhaps) so as to decide which method for learning from disagreement is more appropriate for a given dataset.

In early work (Uma et al., 2021b), we considered a number of approaches to identify the type of disagreement that was most typical in a dataset. However, the objective of the measures used in that work is to identify the type of disagreement in a dataset prior to training a model. In this paper, we report on an investigation of the use of an approach inspired by the idea of **temperature scaling** developed by, e.g., Platt (1999) and Guo et al. (2017) to allow a model to *automatically* adapt in the presence of disagreement in the data. We use a range of datasets known to contain disagreements arising from different sources (Uma et al., 2021b) to train models using the state-of-the-art **soft-loss** approach for learning from disagreement (Peterson et al., 2019; Uma et al., 2020, 2021b) and test whether adding automatic temperature scaling improves model performance. We find that the datasets used can be divided into three groups on the basis of the results obtained with the proposed approach. Automatic temperature scaling works well with datasets in which disagreement is mostly due to substantial overlap between the labels such that annotators have to choose a label more or less randomly. By contrast, the approach does not work at all with data in which the disagreements are due to a clear bias, as in, e.g., subjective tasks such as labeling an item as offensive or not, which is known to be affected by the annotators' political views. Finally, with datasets where most or part of the disagreement arises from linguistic ambiguity lie in between these extremes, suggesting that ambiguity may not sit perfectly within a binary distinction such as the distinction between bias and noise proposed by Reidsma and Carletta (2008).

2. METHODOLOGY: TEMPERATURE-SCALED SOFT LOSS

In this section, we introduce the **temperature-scaled soft loss** approach, which combines the soft loss approach to learning from disagreement we developed in previous work with our own approach to adding temperature scaling in a deep learning model, which we call **automatic temperature scaling**. We first review the soft-loss approach proposed by Peterson et al. (2019) and Uma et al. (2020) and extend soft-loss by including exploration of the suitability of various standard loss functions for soft-loss training. Next, we discuss the (automatic) temperature-scaled soft-loss methodology which involves weighting the soft loss for each item by a learned temperature parameter.

2.1. Soft Loss Learning

The soft-loss functions approach to training from data containing disagreement combines using a standard loss function with a probabilistic soft label generated from crowd annotations (Peterson et al., 2019; Uma et al., 2020). To train a model using the soft-loss function approach, a standard loss function such as cross-entropy or squared error is used; but instead of targeting the ground truth viewed as a one-hot label, a **soft label**—a probability distribution over the labels—is generated from the distribution of crowd labels and used as a target for training the machine learning model. We discuss each step in turn.

2.1.1. Generating Probabilistic Soft Labels

While experimenting with a variety of datasets standardly used for learning from disagreement, Uma et al. (2020) showed that for a soft-loss function, the quality of the predictions is dependent on the method used in generating the probabilistic soft labels, which in turn is dependent on the characteristics of the annotation for the dataset. They evaluated two standard label generation functions—the softmax function and the standard normalization function—finding which is best depends on the dataset. Soft labels obtained through standard normalization were found to be preferable for datasets like CIFAR-10H (Peterson et al., 2019), which were annotated by a large number of expert annotations with high observed agreement among them. Soft labels produced using softmax proved instead more suitable for datasets that do not meet these criteria, such as Gimpel et al.'s POS dataset (Plank et al., 2014a) and the LABELME dataset (Rodrigues and Pereira, 2017). Uma et al. (2021b) further showed that the best soft label for mixed quality datasets, such as PDIS (Poesio et al., 2019),

were obtained by using the posterior distribution of a probabilistic aggregation model such as MACE (Hovy et al., 2013). For our novel misogyny dataset ARMIS (Almanea and Poesio, 2022), we found that the normalized distribution of the annotators was the best-performing label.

2.1.2. A Suitable Loss Function

Peterson et al. (2019) only used the cross-entropy loss function, hypothesizing that it was uniquely suitable for the task. Uma et al. (2021b) tested a variety of other loss functions, including Kullback-Leibler (henceforth: KL) and (Summed) Squared Error (henceforth: SE)[1]. Malinin and Gales (2019) argued that for datasets with high noise due to overlapping labels and resulting in a multi-modal label distribution[2] reverse KL-divergence is most appropriate if the goal is to maximize prediction accuracy. They tested their hypothesis on synthetic data, comparing reverse KL-divergence as a loss function with (forward) KL divergence, and showed that while KL-divergence is a sensible loss function for datasets with low data uncertainty and target distributions where "correct" labels are available, reverse KL-divergence is more suitable when this is not the case.

Thus, as a preliminary experiment, we tested the hypothesis of Malinin and Gales (2019) with our (non-artificial) data by training soft-loss functions for each task using the best soft label and each of the divergence functions. We additionally tested the other two well-known probability-comparing loss functions—the cross-entropy loss function (CE) already used in Peterson et al. (2019) and Uma et al. (2020, 2021b) and the Squared error function (SE) used in Uma et al. (2021b). Soft-loss functions using each of the stated functions can be expressed using the simplified notation:

[1] After some experimentation, and in keeping with the other loss function, we decided to use the sum of the squared errors as opposed to the mean.

[2] Malinin and Gales (2019) use the term **data uncertainty** for this type of noise, but as far as we know their notion of data uncertainty is the same as what Reidsma and Carletta (2008) call random noise.

- Cross-Entropy Soft loss:

$$CE(y_{hum},\ y_\theta) = -\sum_{i=1}^{n} y^i_{hum} \log y^i_\theta \tag{1}$$

- KL Soft loss:

$$D_{KL}(y_{hum} \mid\mid y_\theta) = -\sum_{i=1}^{n} y_{hum} \log(\frac{y^i_\theta}{y^i_{hum}}) \tag{2}$$

- Reverse KL Soft loss[3]:

$$D_{RKL}(y_\theta \mid\mid y_{hum}) = \sum_{i=1}^{n} y^i_\theta \log(\frac{y^i_{hum}}{y^i_\theta}) \tag{3}$$

- SE Soft loss:

$$MSE(y_{hum}, y_\theta) = \sum_{i=1}^{n} (y^i_{hum} - y^i_\theta)^2 \tag{4}$$

where y^i_{hum} is the target label for an item i, the *best soft label*; y^i_θ is the model's predicted probability distribution for that item; and n is the number of items in the training set.

We experiment with these variations of the soft loss function and note the prediction accuracy of the trained models, especially in reaction to Malinin and Gales's (2019) hypothesis. The best soft loss function is used for experiments in automatic temperature scaling.

2.2. Item Weighting Through Automatic Temperature Scaling

One of the most widely adopted approaches to learning from disagreement involves developing methods for identifying **difficult** items–items on which there is an unexpected degree of disagreement among annotators. Such methods typically use statistical inference to infer the difficulty of an item, and then use such difficulty to weigh or filter items classified as intrinsically difficult (refer to, e.g., Carpenter, 2008; Beigman and Beigman Klebanov, 2009; Whitehill et al., 2009) and the discussion of item difficulty approaches in Paun et al. (2022). In the deep learning literature, a number of methods of this type were developed, for which the term **temperature scaling** is often used.

In this paper, we introduce a method of this type, which we called **automatic temperature scaling**, and combine ideas from both temperature scaling and **Platt scaling**. Platt scaling was proposed as a way to calibrate a logistic regression model, i.e., adjust its parameters to reflect uncertainty (Platt, 1999). To calibrate a model, Platt proposes that two **scalar parameters**, a and b ∈ R, be learned by optimizing the negative log-likelihood function over the validation set while keeping the model's parameters fixed. The learned parameters are used to rescale the logits of the model, $\mathbf{z}_i$ resulting in outputs, $f(\mathbf{x}_i) = \sigma(a\mathbf{z}_i + b)$.

[3]In reverse KL, the target human-derived soft label and the predicted soft label are swapped.

Temperature scaling is a single parameter variant of Platt scaling (Guo et al., 2017), where a single scalar parameter, T, called the **temperature**, is used to rescale logit scores for all the classes, $\mathbf{z}_i$, before applying the softmax function. This way, the model's recalibrated probabilities are given as:

$$f(\mathbf{x}_i) = \sigma(\mathbf{z}_i/T) \tag{5}$$

where $\sigma(\cdot)$ is the softmax function. When $T > 1$, the entropy of the output probabilities increases, hence "softening the softmax" and evening out the probability distribution. $T < 1$ hardens the softmax, resulting in a peakier (more modal) probability distribution. Finally, $T = 1$ recovers the unscaled probabilities (Guo et al., 2017). The value of T is obtained by minimizing the negative log-likelihood on a held-out validation dataset. Because T is independent of the class, j, and the item, i, *temperature scaling does not affect which class is predicted and hence does not affect prediction accuracy*.

Automatic temperature scaling, which we propose here, is a natural extension of temperature scaling. It differs from standard temperature scaling in three key ways. First, automatic temperature scaling learns a parameter *vector* T_i jointly as it learns to predict the classes. It does this by learning a network of weights $\mathbf{w}_{T_i}$ and biases b_{T_i} such that

$$T_i = softplus(\mathbf{W}_{T_i}\mathbf{x}_i + b_{T_i}) \tag{6}$$

This network of weights is disjoint from the network of weights for learning to map inputs to targets. By using Softplus as the squashing the function (as opposed to sigmoid, ReLu, or Tanh) we apply non-linearity to the network without overly limiting the bounds of T_i[4].

The reason for moving from a single scalar parameter to a vectorial parameter, and from a single value for the whole corpus to an item dependent parameter, is that difficulty is very much item dependent—e.g., not all images are equally easy or difficult—and also class dependent: some classes are more easily confused than others, as discussed in more detail in the next section. The vectorial expression of temperature is similar to the one used in **matrix scaling**, an alternative temperature scaling also proposed by Guo et al. (2017)[5]. But unlike in matrix scaling (or Platt scaling, in which more than one parameter is also learned), the parameters are not tuned on a held-out validation set; rather, the model jointly learns classifier and scaling parameters. During training, the model's outputs, $\hat{y}_i = f(x_i)$ are computed as follows:

$$f(\mathbf{x}_i) = \sigma(\mathbf{z}_i * T_i) \tag{7}$$

The model's loss is computed using the appropriate soft loss function.

The second key difference is practical in nature but has notable implications. Unlike in temperature scaling, where the logits are divided by temperature T, in automatic temperature scaling,

[4]Sigmoid, ReLu, and Tanh outputs are bounded between [0, 1], [0,1], and [–1, 1] respectively, while Softplus outputs are only lower bounded are zero.

[5]Guo et al. (2017) propose the use of the $max(\cdot)$ function, rather than $softplus(\cdot)$.

the logits are *multiplied* by the temperature; we found this to work better in practice. The consequence is that in automatic temperature scaling, a warmer temperature (higher values of T_i) indicates *lower* uncertainty resulting in peakier probabilities, while colder temperatures indicate higher uncertainty resulting in a more even distribution–the opposite to temperature scaling[6].

The third key difference can be observed from the definition of T_i in Equation (6). Unlike in standard temperature scaling, in automatic temperature scaling, the model does not have a single temperature value; rather, the temperature of any given item is a function of the input vector for the item and the temperature weights of the model, W_{T_i}—the logits for each instance are scaled to a different temperature, determined by the model and learned as a function of the input features of the instance. In this way, if the model is able to identify uncertainty for an input item, it will respond by producing a lower temperature value for that item. The converse is also true. Thus, by considering each instance separately, the model is able to produce temperature values depending on how much data uncertainty it perceives for each item.

This third aspect is vital to understanding the anticipated improvement in predictive accuracy using automatic temperature scaling. In datasets with overlapping labels, because the modal class for affected items is arbitrary, models (much like annotators) are likely to disagree with the modal class of the target labels, predicting a different (and possibly equally plausible) modal class for perceived noisy inputs. The temperature lowering for such items results in a flatter predicted probability distribution and has the added effect of decreasing the loss contribution of that item to the overall loss. Consequently, the model penalizes itself less for such items and reduces the loss contribution of the item to the total loss. In this way, automatic temperature scaling can be comparable to cost-sensitive loss (Plank et al., 2014a).

3. THE EXPERIMENTS

In this section, we present our experimental design and discuss the datasets and models used for the experiments conducted in this study.

3.1. Experiment Design

We conducted the experiments in two phases. First, we experimentally compared the suitability of various standard loss functions for soft loss training as outlined in Section 2.1.2 on several tasks. Then, we extended the best-performing loss function into an automatic temperature-scaled soft loss. For both experiments, we evaluated the models using two evaluation metrics, one hard and one soft.

3.1.1. Hard Evaluation

As a hard evaluation metric, we used accuracy, as done by Peterson et al. (2019) and Uma et al. (2020). We calculated the accuracy of each model's prediction with respect to a standard: the majority vote aggregate of the expert annotators for ARMIS [7] and gold labels for the other datasets.

3.1.2. Soft Evaluation

As noted in previous work (Dumitrache et al., 2018; Peterson et al., 2019; Uma et al., 2020; Basile et al., 2021; Uma et al., 2021b), as the realization that gold labels are an idealization growth, so does the awareness that hard evaluation is not sufficient to compare machine learning models on tasks in which disagreements are extensive, and extremely questionable for tasks in which the labels are subjective and therefore it does not make sense a "gold label" exists that the disagreements can be reconciled to. A particularly obvious illustration of this last point is the **misogyny detection** task, related to hate speech detection. In this task, the labels assigned by annotators are very much dependent on their background, i.e., text found misogynistic by a female annotator or a more liberal annotator may not be found misogynistic by a male annotator or an annotator from a more conservative background.

When evaluating tasks containing disagreements, or in which disagreements may be intrinsic, it would seem insightful not to evaluate models against a questionable gold label only, but also against **soft labels** in the sense discussed above (probability distributions over the labels derived from crowd annotations) in which disagreements are preserved. Consequently, in this paper, our models are also evaluated using a soft evaluation metric, cross-entropy. Like Peterson et al. and Uma et al., we compute the cross-entropy between the probability distribution produced by each model and the **best soft label** produced from the crowd distribution (The label that is most appropriate for that dataset, as discussed above). This form of evaluation provides insight into how well the models are able to capture possible disagreements in labeling resulting from the crowd.

3.2. Data

We used in this study four disagreement-preserving datasets that have been previously used in research into learning to classify from disagreement (Jamison and Gurevych, 2015; Plank et al., 2014a,b; Uma et al., 2020, 2021a; Fornaciari et al., 2021) and that exemplify different sources of disagreement (An in-depth analysis of the disagreements in these datasets has been carried out by Uma et al., 2021b). In addition, we used an entirely new dataset, ARMIS (Almanea and Poesio, 2022), illustrating a different type of disagreement not considered by Uma et al. (2021b): disagreement due to subjectivity.

3.2.1. The Gimpel et al. pos Corpus

The first example of a corpus containing disagreements due to ambiguity (Plank et al., 2014b) is Gimpel et al.'s (2011) POS dataset (henceforth, POS), which has been often used in research into developing disagreement-aware NLP models (Plank et al., 2014a; Jamison and Gurevych, 2015; Fornaciari et al., 2021; Uma et al., 2021b). The dataset consists of 14k Twitter posts annotated with ground truth POS tags collected by Gimpel et al. (2011) from

[6] As such, it would be more appropriate to name T_i "confidence" or "certainty"—but we will stick with the original name to acknowledge the intellectual debt of our proposal to temperature scaling.

[7] Ties were broken by making a random selection.

expert annotators and crowdsourced tags collected by Plank et al. (2014b)—at least five crowdsourced labels per token from 177 annotators.

The workers annotating this corpus often disagree with the ground truth label; the observed agreement (A_o, Artstein and Poesio, 2008) for the dataset is 0.73, as computed using the multi-annotator version of Fleiss Kappa (Fleiss et al., 2004).

A typical example of the disagreements found in this corpus is shown below (the token to be tagged is in bold):

(8)

Noam	likes	**social**	media
Noun	Verb	Adj/Noun	Noun

in the context, the category *Noun* would seem to be just as appropriate as the category *Adj* for the token ***social***.

Plank et al. (2014b) conducted an analysis of the easy and hard cases in this dataset, finding that the vast majority of inter-annotator disagreements are due to **genuine linguistic ambiguity**, as in this example, although the POS categories Adj and Noun are clearly distinct, in some cases, it is not possible to tell what is the "right" category (Plank et al., 2014b). In fact, an analysis of the POS dataset carried out by Uma et al. (2021b) showed that the average observed agreement on an "easy" category such as nouns (particularly for name tokens like Twitter handles) is much higher than for other categories.

For experiments using this dataset, we split the 14k tokens into training (12k) and testing (2k) and use the development dataset released by Plank et al. (2014a) for validation.

3.2.2. The PDIS Corpus

The second corpus we used contains disagreements in part due to ambiguity, in part to annotator carelessness. The *Phrase Detectives* 2 corpus (Poesio et al., 2019) is a crowdsourced anaphoric reference corpus collected with the *Phrase Detectives* game-with-a-purpose (Poesio et al., 2013)[8]. Anaphoric reference is another aspect of linguistic interpretation in which ambiguity is rife (Poesio et al., 2006; Versley, 2008; Recasens et al., 2011). For example, Poesio et al. (2006) discussed examples such as (3.2.2).

(9)

```
3.1    M: can we .. kindly hook up
3.2     : uh
3.3     : engine E2 to the boxcar at ..
          Elmira
4.1    S: ok
5.1    M: +and+ send \textcolor{red}
          {\textbf{it}} to Corning
5.2     : as soon as possible please
6.1    S: okay
          [2sec]
7.1    M: do let me know when it gets
          there
8.1    S: okay it'll /
8.2     : it should get there at 2 AM
9.1    M: great
9.2     : uh can you give the
9.3     : manager at Corning instructions
          that
9.4     : as soon as it arrives
9.5     : it should be filled with
          oranges
10.1   S: okay
10.2    : then we can get that filled
```

In this exchange, it is not clear whether the pronoun *it* in 5.1 (in red) refers to *the engine E2* that has been hooked up to *the boxcar at Elmira* or to the boxcar itself or indeed whether the distinction matters at all. It is only at utterance 9.5 that we get evidence that *it* probably refers to *the boxcar at Elmira* since only boxcars can be filled with oranges. The two interpretations are clearly distinct–the pronoun cannot refer to both–but it is not possible to decide which is the intended one from the context.

The *Phrase Detectives* 2 corpus consists of 542 documents, for a total of 408K tokens and 107K markables, annotated by slightly less than 2,000 players producing a total of 2.2M judgments—about 20 judgments per markable on average. In total, 64.3% of the markables received more than one distinct interpretation from the players. Some of the disagreements are due to annotator error/carelessness, others to interface issues; but for about 10% of markables, disagreement is again due to **genuine linguistic ambiguity**.

In this study, we used PDIS, a simplified version of the corpus containing only binary information status labels: discourse new (DN) (the entity referred to has never been mentioned before) and discourse old (DO) (it has been mentioned). PDIS still consists of 542 documents, for a total of 408K tokens and over 96K markables; an average of 11.87 annotations per markable are preserved[9].

Forty-five of the documents (5.2K markables), collectively called PD_{gold}, additionally contain expert-adjudicated gold labels. This subset of PDIS was designated as the test set. The training and development datasets consist of 473 documents (and 86.9K markables) and 24 documents (4.2K markables), respectively[10].

3.2.3. The LabelMe Corpus

The most widely used corpus for learning to classify images from crowds is the LabelMe dataset[11] (Russell et al., 2008). It classifies outdoor images according to 8 categories: *highway, inside city, tall building, street, forest, coast, mountain*, or *open country*. Using Amazon Mechanical Turk, Rodrigues and Pereira (2017) collected an average of 2.5 annotations per image from 59 annotators for 10K images in this dataset.

The observed agreement for this dataset, also computed using the multi-annotator version of Fleiss et al.'s (2004) Kappa, is 0.73, which is the same level of average observed agreement seen in the POS dataset. However, it can be argued that the source and nature of the disagreement in this dataset are different, consider **Figure 1** for an illustration. The ground truth label for the example image is *inside city*, and one annotator chose that label as well, but two other annotators chose *tall building*. Notice the difference from the ambiguity cases in POS and PDIS: there, two interpretations are possible, but a word can only have one—it is just that it is

[8]https://github.com/dali-ambiguity

[9]DO judgments with different antecedents are considered identical, and the judgments other than DN or DO are removed.

[10]Another example of corpus were the disagreement is due to linguistic ambiguity is Dumitrache et al. (2019).

[11]http://labelme.csail.mit.edu/Release3.0

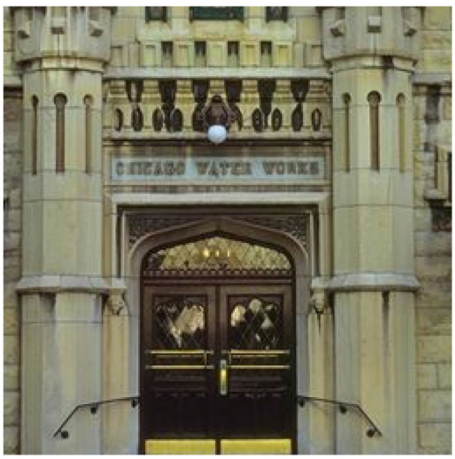

FIGURE 1 | An example of disagreement from LabelMe Ground truth label: *insidecity*, crowd annotations: [*insidecity*:1, *tallbuilding*:2].

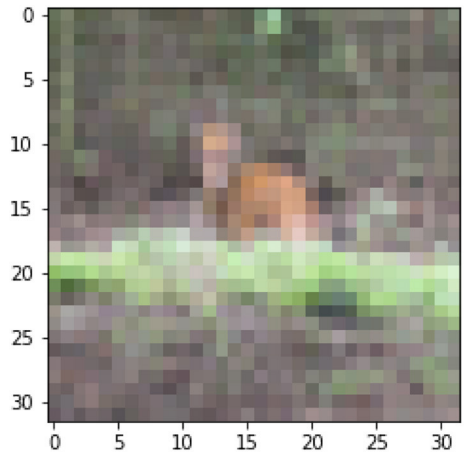

FIGURE 2 | An example of disagreement from CIFAR10H Ground truth label: *deer*, crowd annotations: [*dog*:33, *deer*:13, *horse*:4].

not possible to know which from the context. Here, *both* labels can be applied at the same time. Uma et al. (2021b) carried out an analysis of this dataset, finding that examples like **Figure 1** are prevalent. That is, the disagreement for this dataset is largely due to an **imprecise annotation scheme** where label categories are not necessarily mutually exclusive but may **overlap**. As a consequence, an annotator forced to choose one among the overlapping categories which apply to a particular image will likely make a random choice.

In our experiments, we randomly split the 10K images into training and test data (8,882 and 1,118 images respectively) to allow for ground truth and probabilistic evaluation. A total of 500 images from the dataset with gold labels were used as a development set.

3.2.4. The CIFAR-10H Corpus

As an example of a crowdsourced corpus containing very little disagreement and that primarily due to item difficulty, we used Krizhevsky's (2009) CIFAR-10H dataset, which consists of 60K tiny images from the web, carefully labeled, and expert-adjudicated to produce a single gold label for each image in one of 10 clearly distinct categories: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship,* and *truck*. Peterson et al. (2019) collected crowd annotations for 10K images from this dataset (the designated test portion) using Amazon Mechanical Turk, creating the CIFAR-10H dataset[12], which we use for our experiments.

The observed agreement for this dataset is 0.92, the highest among all the datasets. Clearly, the 2,457 annotators (about 51 annotators per item) found the annotation scheme to be clear and mostly agree with the expert opinion on what the label for each item would be. Notice that unlike in LABELME, there is no overlap: it is not possible for an object to belong to multiple categories. Cases of disagreement among annotators do occur, but they are primarily of the kind illustrated by **Figure 2**, which is because of the poor quality of the image, it is not possible to decide from the picture which animal is illustrated. Yet, there is no question that only one category can apply. We consider such cases as proper examples of **difficult to classify** items—items to which only one category from the scheme applies, yet problematic to classify because of noise.

We used the CIFAR-10H dataset for training and testing using a 70:30 random split, ensuring that the number of images per class remained balanced as in the original dataset. We also use a subset of Krizhevsky (2009) CIFAR-10 training dataset (3k images) as our development set.

3.2.5. The ArMIS Corpus

Finally, to exemplify an important source of disagreement—the fact that certain judgments are intrinsically subjective—we used our own ARMIS corpus (Almanea and Poesio, 2022). ARMIS is an Arabic misogyny dataset. It consists of 1K tweets each annotated with binary labels: 1 if the tweet expresses a misogynistic behavior according to the annotator's subjective point of view, 0 if the annotator believes that the tweet is not misogynistic. The tweets were collected using the Twitter API in October 2020, using a keywords list which was manually created specifically for this task, including specific slang words, phrases, and hashtags in order to get the related tweets, such as "Feminist," "Deficient mind and religion." The important aspect of this dataset is that it was annotated by three experts" annotators, carefully chosen to reflect different political views: liberal, moderate, and conservative. The annotators were asked to annotate the tweets based on their perspective.

The observed agreement of the annotators is 0.77, higher than the observed agreement of both POS and LABELME datasets (0.73), lower than the 0.92 observed agreement of CIFAR-10H, and equal to the observed agreement of PDIS. It is important to note while PDIS and ARMIS have the same level of disagreement, the nature and source of the disagreement for the ARMIS dataset

[12]https://github.com/jcpeterson/cifar-10h

differs from that of PDIS and indeed from the others. While Uma et al. (2021b) show that PDIS disagreements can be attributed to noise from spammers, the ambiguity of labels, or interface problems, an analysis of the disagreement in ARMIS showed the nature of the disagreements to be largely due to the **subjective viewpoints** of the diverse annotators.

For these experiments, we split the 964 tweets in ARMIS into 674 for training, 145 for validation, and 145 for testing. Gold labels were not obtained, as is fitting for a task of such as divisive nature, where annotator background plays a substantial role in how they label. However, as a compromise, we use majority voting to produce a hard label for hard evaluation purposes.

3.3. Base Models

The base models used in these experiments are the state-of-the-art or near state-of-art models used in previous work (Uma et al., 2020; Almanea and Poesio, 2022), many of which were made available to the participants to the 2021 SEMEVAL shared task on learning from disagreement (Uma et al., 2021a). We briefly summarize these models in this subsection.

3.3.1. The POS Tagging Model

For POS tagging, we used the bi-LSTM model (Plank et al., 2016) used by Uma et al. (2020). The model we used is improved from Plank et al. (2016) by using attention over the input token and character embeddings to learn contextualized token representations.

3.3.2. The PDIS Information Status Model

The model for this task was also developed by Uma et al. (2021a). Uma et al. combined the mention representation component of Lee et al.'s (2018) coreference resolution system with the mention sorting and non-syntactic feature extraction components of the IS classification model proposed by Hou (2016)[13] to create a novel IS classification model that outperforms (Hou, 2016) on the PDIS corpus. The training parameters were set following Lee et al. (2018).

3.3.3. The LabelMe Image Classification Model

For the LabelMe image classification, we replicated the model from Rodrigues and Pereira (2017). The images were encoded using pre-trained CNN layers of the VGG-16 deep neural network (Simonyan et al., 2013) and passed to a feed-forward neural network layer with a RELU activated hidden layer with 128 units. A 0.2 dropout is applied to this learned representation which is then passed through a final layer with softmax activation to produce the model's predictions.

3.3.4. The CIFAR-10H-10 Image Classification Model

The trained model provided for this task is the ResNet-34A model (He et al., 2016), one of the best performing systems for the CIFAR-10 image classification. The publicly available Pytorch implementation of this ResNet model was used[14].

[13]This model was developed for fine-grained information status classification on the ISNOTES corpus (Markert et al., 2012; Hou et al., 2013).

[14]https://github.com/KellerJordan/ResNet-PyTorch-CIFAR10

TABLE 1 | The effect of different loss functions for soft loss training on accuracy.

	POS	PDIS	LABELME	CIFAR-10H	ArMIS
SE Soft loss	79.20	92.90	84.21	63.49	76.83
CE Soft loss	79.80	92.86	84.66	66.54	**77.79**
KL Soft loss	**79.96**	92.86	84.73	**66.58**	76.41
Reverse KL Soft loss	79.81	**92.95**	**84.92**	63.71	75.59

The bold values indicate the best results for each model (indicated in the first column) using a given metric (indicated in the column header).

TABLE 2 | Results showing the accuracy (higher is better) and cross-entropy (lower is better) of soft loss models with and without temperature.

Task	Model	Accuracy↑	Cross-entropy ↓
LABELME	Reverse KL soft loss	84.97	1.671
LABELME	Reverse KL soft loss + T_i	**86.29***	**1.656**
POS	KL soft loss	79.96	**1.268***
POS	KL soft loss + T_i	**80.01**	1.547
PDIS	Reverse kl soft loss	92.95	0.467
PDIS	Reverse kl soft loss + T_i	**93.00**	**0.395***
CIFAR-10H	KL soft loss	**66.58***	**1.109***
CIFAR-10H	KL soft loss + T_i	63.89	1.223
ArMIS	CE soft loss	**77.79**	**0.586***
ArMIS	CE soft loss + T_i	76.83	0.636

An asterisk is used to indicate significantly better results.
The bold values indicate the best results for each model (indicated in the first column) using a given metric (indicated in the column header).

3.3.5. The ArMIS Arabic Misogyny Classification Model

For this task and dataset, we fine-tuned the state-of-the-art AraBERT base model (Antoun et al., 2020) with a maximum sequence length of 128, learning rate of 1e-5, batch size of 8, and training for 10 epochs.

4. RESULTS

Table 1 compares the effectiveness of different probability-comparing loss functions for making gold predictions, identifying the best soft loss function for each dataset. **Table 2** presents the results obtained for each task by models using the best soft loss function from **Table 1** with and without automatic temperature scaling, evaluated using both hard and soft metrics.

To account for non-deterministic model training effects, each model was trained and tested several times: (i) 30 times each for POS and LABELME (ii) 10 times each for PDIS, CIFAR-10H, and ARMIS owing to the complexity of the base models. We measure significance *via* bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014). The rest of this section discusses the results from these tables, highlighting significant results. The best result for each dataset is highlighted in bold.

4.1. Choosing the Loss Function

The aim of this preliminary experiment was to investigate Malinin and Gales's (2019) hypothesis that Reverse KL divergence is the most appropriate loss function for training models on

datasets with high data uncertainty. We found that the Reverse KL soft loss function outperforms the other soft loss functions by a noticeable margin (0.19) for one dataset only, LABELME—though this margin is not significant[15]. This is the dataset for which we observe the most disagreement due to an annotation scheme with overlapping labels, as opposed to linguistic ambiguity (as in POS), or a combination of linguistic ambiguity and random noise (as in PDIS), or item difficulty (as in CIFAR-10H), or annotator biases (as in ARMIS). For CIFAR-10H, the dataset with the least amount of disagreement (and noise), as discussed by Uma et al. (2021b), we observe that Reverse KL soft loss falls nearly 3 significance points below either CE or KL soft loss. The SE loss function also performs poorly on this dataset, likely because SE optimizes the loss for non-modal classes, and this is an undesirable trait for a dataset like CIFAR-10H where the modal class is usually the gold class.

Following this experiment, we determine the best soft loss function for each dataset to be used as the starting point for the automatic temperature-scaled soft loss is as follows: CE for ARMIS, KL for POS and CIFAR-10H, and reverse KL for PDIS and LABELME.

4.2. Temperature Scaling Soft-Loss Learning

The first observation emerging from **Table 2** is that automatic temperature scaling only significantly improves results in one task: LABELME. In other words, our results would suggest that automatic temperature scaling only works when a disagreement arises from overlapping labels, resulting in the arbitrariness of ground truth.

In the next two datasets, POS and PDIS, the effect of temperature scaling on the performance of the models are mediocre or non-existent. These are the datasets for which we and Plank et al. (2014b) and Poesio et al. (2019) have shown that although a certain amount of noise is present, the disagreements are largely due to linguistic ambiguity and/or interface limitations.

At the other extreme, we have two datasets in which temperature scaling hurts performance. One of these is CIFAR-10H. This is a dataset with a very high observed agreement, 0.92. We also showed that the very few disagreements in this dataset are due to difficulty experienced by annotators when labeling blurry images. In other words, these disagreements are not systematic or a result of an imprecise annotation scheme but are due to the characteristics of the input. The other dataset for which automatic temperature scaling leads to a reduction in model performance is ARMIS. In this case, there is lower agreement than in CIFAR-10H, but this is not a reflection of systematic noise or data uncertainty, but of annotator uncertainty due to subjective biases.

5. INTERPRETING T_i

Our results show that among the datasets we considered in this study, automatic temperature scaling is effective for the one dataset in which disagreements are primarily due to what we may call **label arbitrariness**: the randomness in judgments originating from the fact that annotators have to choose one between multiple labels all of which could apply to an image and do so without appealing to any theory (given the vagueness of the annotation scheme). In this section, we examine the temperature predictions of the model for this dataset to understand what the model learns about label arbitrariness.

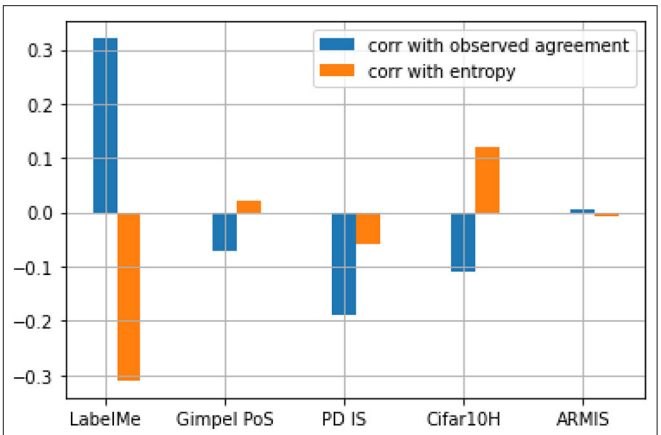

FIGURE 3 | Graph showing the correlation of T_i with observed agreement and entropy.

One way to do this is to measure the correlation of the temperature values to known measures of item agreement/uncertainty/difficulty. **Figure 3** shows the Pearson correlation (Pearson, 1896) between the temperature parameter and two such metrics of uncertainty/difficulty: observed agreement and normalized entropy. The results show that for LABELME, the only dataset for which our method produces a significant improvement over the soft-loss baseline, the model's T_i predictions have the strongest positive correlation to the observed agreement. This means that the model tended to make higher T_i predictions for items with a high observed agreement and lower T_i predictions for items with a low observed agreement. The model also has the strongest negative correlation to entropy. These two results suggest that for this dataset (but not for others), T_i is a moderately good predictor of uncertainty for this dataset as measured by observed agreement and entropy. What is it about the type of disagreement due to annotation schemes in which labels overlap that explains why temperature scaling improves performance with this kind of dataset, but not with others?

As mentioned earlier, in the one study of the differences between types of disagreement we are aware of, Reidsma and Carletta (2008) proposed a distinction between two types of disagreement between annotators and argued that they affect the performance of machine learning models in different ways. One kind is disagreements due to **random noise**, not conforming to any theorizable pattern. A second type is disagreements due to **bias**, which are identifiable through the occurrence of patterns of disagreement. The fact that automatic temperature scaling works best for disagreement due to overlap, which is the type of disagreement among those we studied that most

[15]Significance was computed using bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014).

resemble random noise because the annotators have to choose randomly; and it works worst for the clearest case of bias among our datasets, the misogyny data might suggest that automatic temperature scaling is a good method for adjusting model weights when the disagreements are due to random noise, but not when disagreement is due to bias. The mediocre results with PDIS and POS suggest that disagreements due to linguistic ambiguity sit somewhere in the middle, or do not fit this distinction at all. Of course, more research is needed to verify if this hypothesis also holds with other datasets in which disagreement is due to noise.

An alternative explanation can be found in the experiments conducted by Malinin and Gales (2019), who posit that overlapping labels (due to imprecise annotation schemes) introduce **data uncertainty**, resulting in multi-modal distributions[16]. The key characteristic of data uncertainty disagreement is that it is fully observable given the inputs and targets, without the need to appeal to linguistic theory (as in linguistic ambiguity) or annotator background (as in subjectivity disagreement). As such, a network of weights and biases (a machine annotator if you will), given the inputs and label distribution would also experience uncertainty predicting the targets for such images as human annotators do. In fact, an examination of the model's output distribution for the instances with the lowest temperature predictions shows that the model assigned the lowest temperatures (= highest uncertainty) to images belonging to the categories *tall building*, *street*, or *inside city*, the categories for which the annotators most disagree with the gold (**Figure 4** shows the class proportions of images 1st quartile range of temperature while **Figure 5** shows the confusion matrix between the majority and the gold). By calibrating its predictions by its level of certainty for each item, the model was able to fine-tune and improve its performance. Again, more research with other datasets characterized by data uncertainty will be required.

6. CONCLUSION AND FUTURE WORK

Not all disagreements are the same, and it has been shown that not all approaches for learning from disagreement work equally well with datasets containing different types of disagreement (Uma et al., 2021b). In this paper, we reported on experiments on the use of automatic temperature scaling in a learning-from-disagreements setting as a way for automatically adjusting a model to take into account the peculiarities of a particular dataset. Our results show that model calibration *via* automatic temperature scaling can be a simple yet effective approach to improving model performance, particularly with learning ground truth predictions, but only with high disagreement datasets where the disagreements are due to overlapping labels.

We analyzed the temperature values of the successful model in a dataset of this type, to find that the temperature values have some correlation with two known measures of item disagreement/uncertainty—a positive correlation of about 0.3 with an observed agreement and a negative correlation of about 0.3 with entropy. We also observed that the model assigns the lowest temperature to instances with one of the three categories *inside city, street, tall building* shown by Uma et al. to be overlapping. We also found, however, that in datasets where disagreement is due to different reasons, the approach does not work so well.

We provide two possible explanations: automatic temperature scaling provides a good model of uncertainty when disagreements are due to random noise, but not when they are due to biases and automatic temperature scaling is a

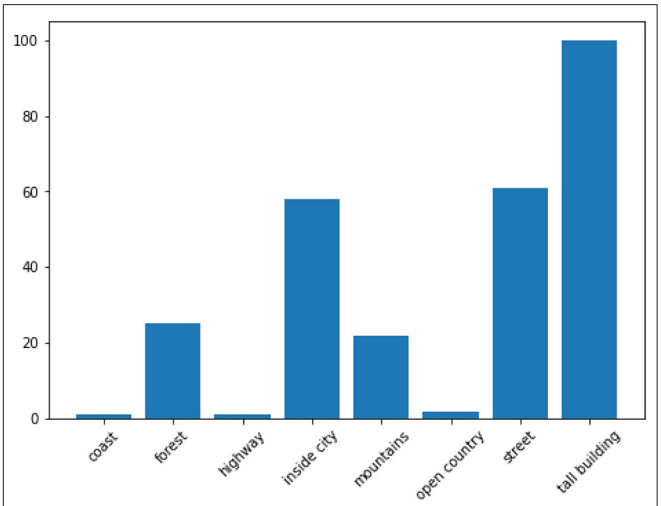

FIGURE 4 | Bar chart showing the gold label distribution of the images with the lowest temperature (images in the 1st quartile range of temperature), i.e., the lowest certainty.

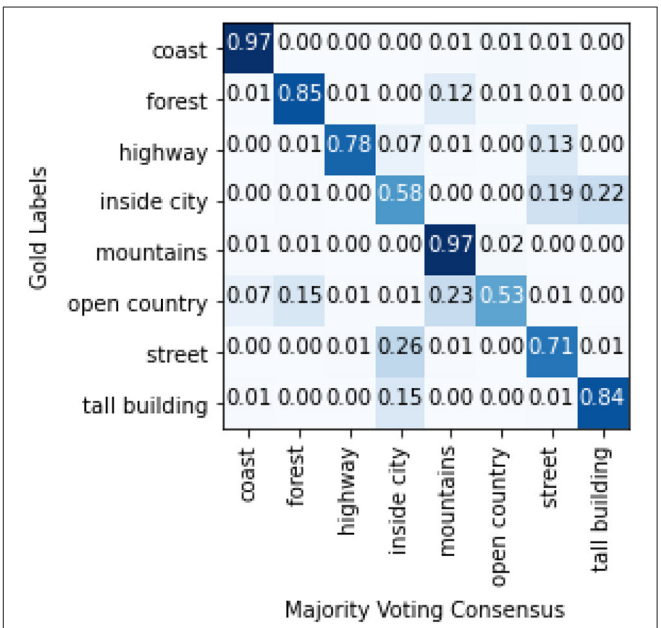

FIGURE 5 | Confusion matrix between gold labels and majority voting consensus for LabelMe.

[16]The results from **Table 1** while not significant do suggest that of the datasets examined in this work, LABELME has the most data uncertainty.

good indicator of data uncertainty. Further research is however needed to test these explanations with other datasets with the same characteristics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://zenodo.org/record/5130737#.YP_V9o5KiUk.

AUTHOR CONTRIBUTIONS

AU: conceptualization, methodology, software, formal analysis, investigation, visualization, and writing. DA: software, investigation, data curation, and writing. MP: conceptualization, methodology, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

AU and MP are supported by the DALI project, ERC Advanced Grant 695662 to MP. DA was supported by a studentship from the Saudi government.

ACKNOWLEDGMENTS

We thank Valerio Basile, Bob Carpenter, Tommaso Fornaciari, Dirk Hovy, Becky Passonneau, Silviu Paun, and Barbara Plank for many useful discussions and comments.

REFERENCES

Almanea, D., and Poesio, M. (2022). The ARMIS dataset of misogyny in arabic tweets. Submitted for publication.

Antoun, W., Baly, F., and Hajj, H. (2020). "AraBERT: transformer-based model for Arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection* (Marseille: European Language Resource Association), 9–15.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.* 34, 555–596. doi: 10.1162/coli.07-034-R2

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., et al. (2021). "We need to consider disagreement in evaluation," in *BPPF* (Online).

Beigman, E., and Beigman Klebanov, B. (2009). "Learning with annotation noise," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Suntec: Association for Computational Linguistics), 280–287.

Beigman-Klebanov, B., and Beigman, E. (2009). From annotator agreement to noise models. *Comput. Linguist.* 35, 495–503. doi: 10.1162/coli.2009.35.4.35402

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). "An empirical investigation of statistical significance in NLP," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island: Association for Computational Linguistics), 995–1005.

Carpenter, B. (2008). *Multilevel Bayesian Models of Categorical Data Annotation*. Available online at: http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf

Dawid, A. P., and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *J. R. Stat. Soc. Ser. C* 28, 20–28.

Dumitrache, A., Aroyo, L., and Welty, C. (2018). "Crowdsourcing semantic label propagation in relation classification," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels: Association for Computational Linguistics), 16–21.

Dumitrache, A., Aroyo, L., and Welty, C. (2019). "A crowdsourced frame disambiguation corpus with ambiguity," in *Proceedings of North American Chapter of the Association for Computational Linguistics* (Minneapolis, MN).

Fleiss, J. L., Levin, B., and Paik, M. C. (2004). *The Measurement of Interrater Agreement*, 598–626. Sydney, NSW: John Wiley & Sons, Ltd. p. 598–626.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). "Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning," in *Proceedings of North American Chapter of the Association for Computational Linguistics* (Online).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al. (2011). "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, OR: Association for Computational Linguistics), 42–47.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW), 1321–1330.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Amsterdam: IEEE), 770–778.

Hou, Y. (2016). "Incremental fine-grained information status classification using attention-based lstms," in *COLING* (Osaka).

Hou, Y., Markert, K., and Strube, M. (2013). "Global inference for bridging anaphora resolution," in *HLT-NAACL* (Seattle, WA).

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, GA: Association for Computational Linguistics), 1120–1130.

Jamison, E., and Gurevych, I. (2015). "Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon: Association for Computational Linguistics), 291–297.

Krizhevsky, A. (2009). *Learning Multiple Layers of Features From Tiny Images*. Available online at: https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Lee, K., He, L., and Zettlemoyer, L. (2018). "Higher-order coreference resolution with coarse-to-fine inference," in *NAACL-HLT* (New Orleans, LA).

Malinin, A., and Gales, M. (2019). "Reverse kl-divergence training of prior networks: improved uncertainty and adversarial robustness," in *NeurIPS* (Vancouver, BC).

Markert, K., Hou, Y., and Strube, M. (2012). "Collective classification for fine-grained information status," in *ACL* (Jeju).

Paun, S., Artstein, R., and Poesio, M. (2022). *Statistical Methods for Annotation Analysis*. Claypool, IN: Morgan.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., and Poesio, M. (2018). Comparing bayesian models of annotation. *Trans. Assoc. Comput. Linguist.* 6, 571–585. doi: 10.1162/tacl_a_00040

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. Ser. A* 187, 253–318.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. (2019). "Human uncertainty makes classification more robust," in *2019 IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 9616–9625.

Plank, B., Hovy, D., and Søgaard, A. (2014a). "Learning part-of-speech taggers with inter-annotator agreement loss," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg: Association for Computational Linguistics), 742–751.

Plank, B., Hovy, D., and Søgaard, A. (2014b). "Linguistically debatable or just plain wrong?," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Baltimore, MD: Association for Computational Linguistics), 507–511.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Berlin: Association for Computational Linguistics), 412–418.

Platt, J. C. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers* (Nebraska, NA: MIT Press), 61–74.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Intell. Interact. Syst.* 3, 1–44. doi: 10.1145/2448116.2448119

Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., and Kruschwitz, U. (2019). "A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN: Association for Computational Linguistics), 1778–1789.

Poesio, M., Sturt, P., Arstein, R., and Filik, R. (2006). Underspecification and anaphora: theoretical issues and preliminary evidence. *Discourse Processes* 42, 157–175. doi: 10.1207/s15326950dp4202_4

Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., and Moy, L. (2010). "Learning from crowds," *J. Mach. Learn. Res.* 11, 1297–1322. Available online at: https://jmlr.org/papers/v11/raykar10a.bib

Recasens, M., Hovy, E., and Martí, M. A. (2011). Identity, non-identity, and near-identity: addressing the complexity of coreference. *Lingua* 121, 1138–1152. doi: 10.1016/j.lingua.2011.02.004

Reidsma, D., and Carletta, J. (2008). Reliability measurement without limits. *Comput. Linguist.* 34, 319–326. doi: 10.1162/coli.2008.34.3.319

Rodrigues, F., and Pereira, F. (2017). Deep learning from crowds. Available online at: https://www.bibsonomy.org/bibtex/1a9d85dd14f242883706b4b37e716429e/ghagerer

Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *Int. J. Comput. Vision* 77, 157–173. doi: 10.1007/s11263-007-0090-8

Sheshadri, A., and Lease, M. (2013). "Square: a benchmark for research on computing crowd consensus," in *HCOMP* (Palm Springs).

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv[Preprint].arXiv:1312.6034.* doi: 10.48550/arXiv.1312.6034

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. (1994). "Inferring ground truth from subjective labelling of venus images," in *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94 (Cambridge, MA: MIT Press), 1085–1092.

Søgaard, A., Johannsen, A., Plank, B., Hovy, D., and Alonso, H. M. (2014). "What's in a p-value in nlp?," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (Baltimore, MD), 1–10.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J. P., Plank, B., et al. (2021a). "Semeval-2021 task 12: learning with disagreements," in *SEMEVAL* (Online).

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2020). "A case for soft-loss functions," in *Proc. of HCOMP* (Online).

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021b). Learning with disagreements. *J. Art. Intell. Res.* 4, 201–213.

Versley, Y. (2008). Vagueness and referential ambiguity in a large-scale annotated corpus. *Res. Lang. Comput.* 6, 333–353. doi: 10.1007/s11168-008-9059-1

Whitehill, J., Fan Wu, T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, BC: Curran Associates, Inc), 2035–2043.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Uma, Almanea and Poesio. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.