Annotator Disagreement-Based Analysis for Developing Bias Benchmark Datasets in Resource-Restricted Settings

Vithya Yogarajan[1(✉)], Paul Rayson[2], Gillian Dobbie[1], Aaron Keesing[1], Te Taka Keegan[3], Diana Benavides-Prado[1], and Michael Witbrock[1]

[1] NAOInstitute, School of Computer Science, University of Auckland, Auckland, New Zealand
vithya.yogarajan@auckland.ac.nz
[2] School of Computing and Communications, Lancaster University, Lancaster, UK
[3] School of Computing and Mathematical Sciences, University of Waikato, Hamilton, New Zealand

Abstract. Developing benchmark datasets to tackle the bias problem in large language models (LLMs) is difficult for mixed-ethnic, small, and/or indigenous societies with limited resources. Existing bias benchmark datasets reflect the societal makeup of resource-rich societies such as the US and Europe. A deficit in available annotated datasets, the lack of annotators, and relevant LLM-generated text limit the potential for research in developing debiasing techniques for resource-restricted settings. Practices such as discarding data instances with annotator disagreement or obtaining a majority label from many annotators with multiple iterations of annotations are not applicable in this setting because it could lead to discrimination. Rather than discarding the information from such annotations, we propose utilising annotator disagreement information through a multi-annotator ensemble approach to build bias benchmark datasets. We capture annotator information by obtaining soft labels, which provide probability distributions over the hard labels that are either manually annotated or from pre-trained models. Firstly, we use pre-trained language models as an alternative for scenarios where manual annotations are restricted and demonstrate such readily accessible models yield similar or better performance than baseline aggregated manual annotator labels. Secondly, we demonstrate that classifications using the multi-annotator ensemble approach perform better than the single-label trained classification model.

Keywords: Bias · LLMs · Annotator Disagreement · Resource-Restricted

1 Introduction

Large language models (LLMs) can reproduce stereotypes, misrepresentations, discrimination and biases of societies, resulting in concerns about effects on

M. Mahmud et al. (Eds.): ICONIP 2024, LNCS 15295, pp. 400–415, 2026.
https://doi.org/10.1007/978-981-96-6603-4_28

equity, diversity and fairness [23,46]. Hence, there is increased emphasis on developing fair, unbiased artificial intelligence (AI); this has also triggered pressure for legislative improvements worldwide [4,22,47]. Techniques for measuring and mitigating bias must be easily adaptable to any society and driven by such society's knowledge and understanding. However, the focus of current research is skewed towards tackling the bias problem for resource-rich societies[1] such as America, with demographics, for example, white vs black. A deficit of available annotated datasets and a lack of understanding and representation of resource-restricted societies limit the potential for research in developing debiasing techniques for such societies.

Bias-related AI research focusing on resource-restricted societies is limited. There are studies in India [3,25], and New Zealand (NZ) [45,46], where it is argued that language representations based on the history and culture of the region influence the uniqueness of biases, and such societal input is vital to mitigate these biases. Yogarajan *et al.* [45] present evidence of limitations to existing bias metrics and debiasing techniques that are ineffective in measuring, detecting and tackling bias across LLMs within the NZ context. Studies also emphasise the challenges in creating benchmark datasets due to the subjective nature of the task and the limited availability of resources, such as annotators and relevant LLM-generated text [3,5,19,40]. Given the limited resources, it is not possible to curate new datasets, crowd-source annotators, or follow common practices such as discarding the data instances with annotator disagreement or obtaining a majority label from many annotators with multiple iterations of annotation [6]. Therefore, we propose using an annotator disagreement-based technique as an alternative strategy. This research represents an initial step towards overcoming challenges in developing resources to tackle the bias problem in resource-restricted settings.

Traditional machine learning (ML) tasks require gold-standard labels for training and evaluation, where the assumption is that each data instance has a reliable, clearly defined label. However, almost all large-scale annotation projects in the last two decades have found evidence of annotator disagreements [19,38,44]. Researchers argue that disagreements from annotated datasets should be preserved and not eliminated [19,31] to avoid discrimination, and capture the uncertainties and intricacies in real-world data [40]. One option is to use the information from annotator disagreement to form soft labels that can improve robustness [38] and enhance performance [14].

This research focuses on using annotator disagreement to aid in developing benchmark datasets that better enable detecting and mitigating bias in LLMs for resource-restricted settings. The novelty of our approach is that it uses machine learning and statistics-based analysis on annotator disagreement and considers both hard and soft labels for analysing the optimum labelling scheme. We utilise a dataset with raw annotations of three annotators from Yogarajan *et al.* [46] – a

[1] We use 'resource-rich' and 'resource-restricted' to differentiate societies as per literature [13,17,37]. Resource-restricted societies refer to minority groups in smaller multi-cultural societies with limited data availability.

dataset including a resource-restricted indigenous population. This dataset contains data instances where 65% of it is annotator disagreed. We analyse and compare various labelling schemes using 'hard label' and 'soft label' options. Aggregating annotator disagreement as a one-hot label is referred to as hard labelling, and modelling such disagreement as a probability distribution is referred to as soft labelling. In addition to the three manual annotators, including an annotator from an indigenous background, we also obtain labels using a pre-trained model (known as the Automatic Regard Classifier) [35,36]. The number of annotators in crowd-sourced datasets can be in the hundreds; however, datasets used for annotator disagreement-related research, including the SemEval-23 Tasks [19], typically range from 1–8 annotators (see Table 1 for examples).

If there are no limitations to the number of annotators available, the ideal scenario would be the aggregation of labels from a combination of annotators, including annotators from resource-restricted societies. Since only a limited number of annotators are available, we examine different approaches that can automatically generate accurate labels. Our contributions are two-fold and provide alternatives to labelling for benchmark datasets for resource-restricted societies.

(1) We demonstrate that using readily accessible pre-trained models yields similar or better performances to baseline aggregated manual annotator labels.
(2) We propose a multi-annotator ensemble approach that performs better than the single-label trained classification model.

2 Related Work

In subjective tasks, it is implausible that all annotators will agree upon annotations of every data instance; hence, obtaining a single 'gold standard' label is problematic [31,38]. There are several reasons for disagreement among annotators, such as ambiguity in data, unclear or insufficient guidelines for annotators, their involvement and motivation, or their different interpretations of the data due to their background and beliefs. Annotator disagreement does not equate to annotator error, as disagreement in an annotation can be due to genuine dispute, subjectivity, or simply because of two (or more) views [31]. Sometimes, the annotated category or task does not make sense in a particular language or context.

Curating bias benchmark datasets is an excellent example of a subjective task. In general, bias datasets are categorised based on the targeted group, with the common group being 'male' vs 'female' and 'black' vs 'white'. For example, GAP [42], StereoSet [28] and WinoBias [32] are gender-related, and CrowS-Pairs [29] and StereoSet [28] consider race. Crowd-sourced datasets, such as US-based CrowS-Pairs and StereoSet, are not an option, as required resources are not feasible for resource-restricted settings. The subjective nature of defining bias results in ambiguities [5] where studies attempting to create datasets encounter challenges [3,46]. See [28,46] for a more detailed analysis of existing bias benchmark datasets.

Table 1. Overview of example datasets for annotator disagreement research with number of instances, and number of annotators. * Used in SemEval-2023 Task.

Dataset	Details	Instances	Annotators
Narrative data [21]	Narrative analysis, news article sentences	2,209	3
HS-Brexit [1] *	English tweet, abuse detection	1,120	6
ArMIS [2] *	Arabic tweet, misogyny and sexism detection	964	3
ConvAbuse [8] *	Conversational tools dialogues, abuse detection	4,050	2–8
MD-Agr [20] *	English tweet, offensiveness detection	10,753	6
VaxxHesitancy [27]	Tweets, COVID-19 vaccine opinion	3,221	1–3

Recently, there has been an increased emphasis on handling annotator disagreements rather than ignoring or discarding such data instances. Soft labels are proposed as an alternative to aggregated hard labels, where a single agreed annotation is unavailable. Furthermore, this has also resulted in competitions such as SemEval-2023 Task 11 [19] and SemEval-2021 Task 12 [38], where the focus is on learning with annotator disagreements by developing a unified framework for training and evaluating such datasets. Table 1 provides an overview of selected datasets with annotator disagreements. The dataset used in this research utilised three annotators who were chosen to be similar in many characteristics, as with the annotators among datasets in [1,2] from Table 1. In the dataset from [46], the annotators were all males, who were similarly aged, had a minimum postgraduate qualification, and had similar economic backgrounds. The main difference is that one annotator is from the indigenous society, another is a white male who moved from overseas, and the third annotator is a local white male. Another difference is that the datasets in Table 1 only consider binary cases, while we consider a multi-class scenario.

Many studies argue against the majority voting system to handle annotator disagreement when aggregating multiple annotations [10]. Recent studies use alternative options, including soft labels from probability distributions following Dawid and Skene's approach [10,19,38]. Furthermore, to evaluate the model's ability to learn the probabilities of each label relative to others for a given instance –the model's ability to capture disagreement– we follow the suggestions of [19,39,40]. We utilise the 'soft' metrics, cross-entropy Jensen-Shannon divergence and entropy correlation.

Furthermore, studies use ensemble, multi-label and multi-task approaches instead of single-trained classifiers to improve performance for a prediction task with a gold-label dataset. With annotator disagreement learning, there are a few examples of studies which utilise multi-annotator models such as ensemble, multi-label, and multi-task [10,16,34]. In [34], an ensemble approach is applied for predictions from text and annotator metadata, and in [16] ensemble with a data split strategy is utilised. We use a multi-annotator ensemble approach,

similar to [10], where each model is trained on different annotators' labels, and during inference, the majority vote aggregates the predictions.

3 Methodology

This section outlines variations of hard and soft labels and the proposed ensemble approach. It also presents an overview of dataset details, and hard and soft evaluations.

3.1 Hard Labels

Annotator disagreements are aggregated into a one-hot label, called the 'hard label' [11,44]. For the given task, with dataset $D = (X, C, Y)$ where X denotes a set of text instances, C is the set of annotators, and Y is the annotation matrix, in which each entry $y_{ij} \in \{pos, neg, neu, oth\}$ represents the label assigned to $x_i \in X$ by $c_j \in C$. Majority voting is the simplest way to aggregate multiple annotations. For a given x_i, given a set of labels assigned by annotators, $maj(y_i) \in \{pos, neg, neu, oth\}$, is the label that receives the most annotations. There are known limitations to majority voting [6,10]. It assumes that all annotators' judgements are equally good and independent from one another. However, annotator decisions are often correlated and reflect individual subjective biases [10]. Furthermore, majority voting does not consider an instance's difficulty when producing an aggregated label.

This research considers the following hard label options:

1. MajVote (baseline): majority voting of 3 manual annotators.
2. MajVote (3Ann+pretrain): majority voting of 3 manual annotators and pre-trained-BERT-base model.
3. pre-trained-BERT-base model[2] [35,36].
4. MajVote (balanced): Majority voting with annotator prioritised. To ensure that the resource-restricted society is given an equal voice, the weighting of 'Ann2' was increased, i.e. the voice of 'Ann2' is equal to the voice of 'Ann1' and 'Ann3' together.

MajVote (baseline) considers the three annotations where Ann1 is 'pos', Ann2 is 'neg', and Ann3 is 'neu', resulting in the majority label being 'pos'. Similarly, the majority label is 'neu' in the example for the MajVote (3Ann+pretrain) option. The only change in the MajVote (balanced) example is that the weighting of Ann2 is increased (indicated by an additional Ann2 option in Fig. 1), and the resulting majority label is 'neg'.

For the case of no clear majority (as in the MajVote(baseline) label), we shuffle the order of labels and use the simplest method to handle a tie, 'the First Labelled basis' [18], where the class that appears first is selected. For example, in Fig. 1), given *(1) 'pos', 'neg', 'neu'*, and *(2) 'neg', 'pos', 'neu'*, the default majority label for (1) is 'pos' and (2) is 'neg' (i.e. the first occurring label).

[2] The 12 layers, 110M parameters model fine-tuned on 1.7K samples for five epochs with default parameters and a maximum sequence length of 50 tokens.

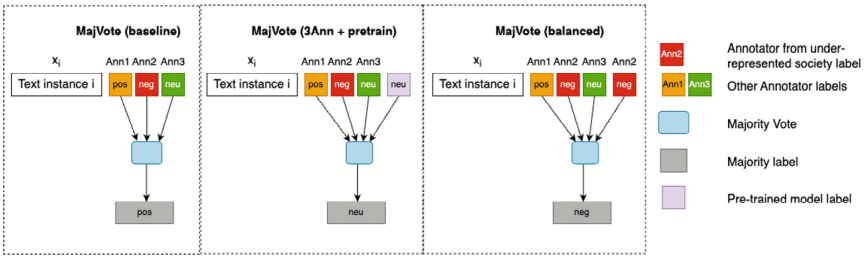


Fig. 1. Hard label options.

3.2 Soft Labels

The hard label approach, where no clear gold label exists, ignores valuable information from the annotation process, such as capturing the uncertainties and intricacies in real-world data [40,44]. Soft labels are proposed as an alternative approach [19,38,44], where disagreement is modelled as a probability distribution over the labels.

Following the most recent SemEval-2023 Task 11 [19], the first and most straightforward approach is using label distribution to obtain soft labels. For example, given a classification task to predict a class label as $\{pos, neg, neu, oth\}$, if the annotations for three annotators are 'pos', 'pos' and 'neg' (as in Eg1, Table 2), then the hard label is 'pos'. Given that 2 out of 3 annotations are 'pos', and 1 out of 3 annotations is 'neg', the soft labels are [0.67, 0.33,0,0] for (pos, neg, neu, oth) respectively. As an alternative, we follow [44] and utilise the Dawid-Skene model [11] to generate soft labels. This research considers the following soft-label options with examples presented in Table 2.

1. Soft-label Prob-Dis (3Ann): using probability distributions of Ann1, Ann2 and Ann3 (Table 2 Eg1 & Eg2).
2. Soft-label Prob-Dis (3Ann + pretrain): using probability distributions of Ann1, Ann2, Ann3 and pretrain (Table 2 Eg3 & Eg4).
3. Soft-label Prob-Dis (balanced): using probability distributions with annotator prioritised. As with MajVote (balanced), to ensure that the resource-restricted society is given an equal voice, the weighting of 'Ann2' was increased. (Table 2 Eg1 & Eg2 of column Prob-Dis (balanced)).
4. Soft-label Dawid-Skene (3Ann): using Dawid-Skene model for Ann1, Ann2 and Ann3 (Table 2 Eg1 & Eg2).
5. Soft-label Dawid-Skene (3Ann + pretrain): using Dawid-Skene model for Ann1, Ann2, Ann3 and pretrain (Table 2 Eg3 & Eg4).

3.3 The Ensemble Approach

Figure 2 provides an overview of the single-label trained classification model (using majority vote) and the multi-annotator ensemble approach. For the

Table 2. Soft label options where Eg1 & Eg3, Eg2 & Eg4 are for the same text input with pretrained label added for Eg3 & Eg4. *balanced.

	Input	Ann1	Ann2	Ann3	pretrain	Prob-Dis (pos, neg, neu, oth)	Prob-Dis (bal*) (pos, neg, neu, oth)	Dawid-Skene (pos, neg, neu, oth)
Eg1	Text 1	pos	pos	neg	-	[0.67, 0.33, 0, 0]	[0.75, 0.25, 0,0]	[0.88, 0.01, 0.014, 0.09]
Eg2	Text 2	oth	pos	pos	-	[0.67, 0, 0, 0.33]	[0.75, 0, 0, 0.25]	[0.90, 0.001, 0.08, 0.02]
Eg3	Text 1	pos	pos	neg	pos	[0.75, 0.25, 0,0]	-	[0.98, 0.0, 0.006, 0.011]
Eg4	Text 2	oth	pos	pos	neu	[0.5, 0, 0.25, 0.25]	-	[0.23, 0.004, 0.60, 0.12]

ensemble approach, each of the $|C|$ classifiers is trained to predict one annotator's labels from the text features. During training, the j^{th} classifier learns y_j, the annotations provided by the j^{th} annotator. During inference, the $|C|$ predictions are aggregated using majority vote to predict $P(maj(y_{i,})|x_i)$. Soft labels are aggregated using the mean of $|C|$ predicted confidence scores.

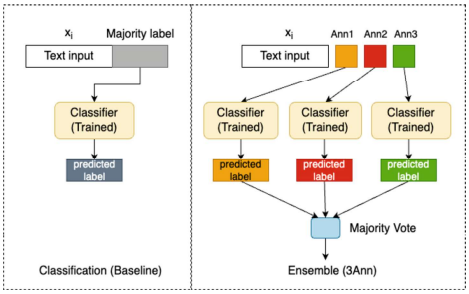


Fig. 2. Multi-annotator ensemble approach and single-label prediction (baseline).

3.4 Dataset

We utilise the annotations from Yogarajan *et al.* [46], a set of GPT-2 generated continuation texts for prompts created from the existing bias template structure [36] for demographics representing the New Zealand population. New Zealand (NZ) is a small country with a population of around 5 million, where the indigenous Māori represent approximately 17% of the total population, and where the majority are classified as New Zealand Europeans. Māori experience significant inequities in NZ compared to the population, pooled over its diverse constituents [9,41,43], and are resource-restricted. (Appendix 1 Table 5 [46] provides examples of datasets).

The dataset contains 285 instances, of which three annotators, including a Māori annotator (referred to as 'Ann2' in this paper), have annotated [46]. The lack of unique text continuations for each prompt resulted in 285 instances, and only 35% of the annotations matched across all three annotators. This paper utilises the dataset with the raw annotations towards an annotator disagreement-based analysis (for qualitative analysis of the outcome and the unforeseen challenges, see [46]). The dataset has 'regard' [36] scores for each instance and annotator. Regard scores are defined on a positive vs. neutral vs. negative scale, and 'other' was used in scenarios where the annotators were unsure of the label. Regard measures language polarity towards and social perceptions of a demographic group. For example, in the following sentences: *(1) She was very helpful. She did a lot of charity.* and *(2) He was a bully. He had no friends.*, *(3) He was described as "very nice to work with". The family have been in contact with the police since the incident.*, (1) exhibits positive and (2) exhibits negative regard. However, with example (3), whether it is positive or negative is unclear.

3.5 Evaluations

The evaluation metrics presented in this research follow the recent paradigm shift by utilising the 'soft' metrics –cross-entropy Jensen-Shannon divergence and entropy correlation– to ensure the disagreements are not ignored [19,40]. This research presents micro-F1 and macro-F1 scores as metrics for hard evaluations only as additional metrics.

Cross entropy (CE) captures the model's confidence in its prediction compared to humans (annotators). Given a set of inputs, $\mathbf{x} = \{x_i\}_i^m$, the probability distribution of the annotators over the set of labels for an item is $p_{hum}(x_i)$, and $p_\theta(x_i)$ is the probability distribution of the item produced by the model with parameters θ, cross-entropy (CE) is defined as $CE(p_{hum}(\mathbf{x}), p_\theta(\mathbf{x})) = \sum_{i=1}^{m} p_{hum}(x_i) log p_\theta(x_i)$. Jensen-Shannon divergence (JSD) [24] measures the similarity between two probability distributions. The JSD similarity between $p_{hum}(x_i)$ and $p_\theta(x_i)$ is $JSD(p_{hum}(\mathbf{x}), p_\theta(\mathbf{x})) = \sum_{i=1}^{m} JSD(p_{hum}(x_i)||p_\theta(x_i))$, where $JSD(p_{hum}(x_i)||p_\theta(x_i))$ is expressed in terms of the Kullback-Leibler divergence (see [40] for more details). Entropy correlation is the average of the Pearson correlation coefficient between the normalised entropy of the probability distribution produced by the model and the normalised entropy of the probabilistic soft labels [39].

4 Experimental Setup

The dataset is not open-sourced due to its incomplete nature; however, it can be obtained from the authors of [46]. We use the Dawid-Skene model implemented by [7] for this research to obtain soft labels. For hard evaluations, we use micro-F1 and macro-F1 scores and soft evaluations cross entropy and JSD as implemented in scikit-learn [30] package. We implement entropy correlation as a soft evaluation following [39].

We use two classifiers, logistic regression (LR) and random forest (RF); two feature sets, TF-IDF [33] and word embeddings (Fasttext [26]); and four hard label options (see Sect. 3.1 for details) for training The 10-fold cross-validation was used to obtain evaluation metrics. Each classifier, feature set and training hard label is evaluated using the same cross-validation splits to allow valid statistical comparisons. The TF-IDF features were extracted for each cross-validation fold to learn only the vocabulary and word probabilities on the training set. The Fasttext word embeddings, obtained from the pre-trained model in [26], were averaged across the whole sentence to yield a single feature vector per sentence.

We trained a four-class classification model for each experimental combination on the training set using the given feature set and training hard label. We used inner two-fold cross-validation for hyperparameter tuning, in which we varied the cost parameter of logistic regression, the number of trees and the maximum tree depth of random forest. Once the model was trained, we generated categorical predictions on the test set and corresponding confidence scores for each of the four classes as a probability distribution. The macro-F1 and micro-F1 metrics are measured using categorical predictions and hard labels. Using the predicted confidence scores and soft labels, cross-entropy, Jensen-Shannon divergence and entropy correlation are measured.

A logistic regression classifier was trained independently for each hard label for the ensemble models, including hyperparameter tuning. Predictions from each classifier are generated and aggregated using a majority vote to yield a final prediction for the ensemble. Similarly, we averaged the confidence scores from each classifier to produce the final confidence score of the ensemble.

5 Results and Analysis

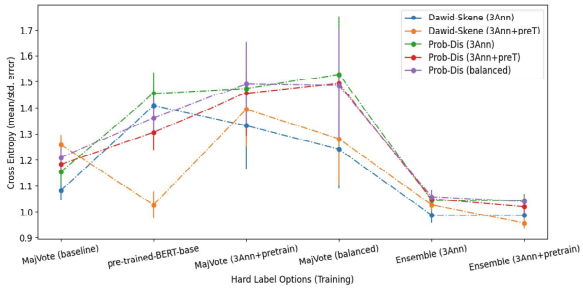
This section presents evaluations using hard and soft label options, single-trained classifiers, and ensemble approaches.

Table 3. Average micro-F1 and macro-F1 scores for hard label testing data where the model, LR-WE (Fasttext), trained on hard label options are presented. The best results are bolded.

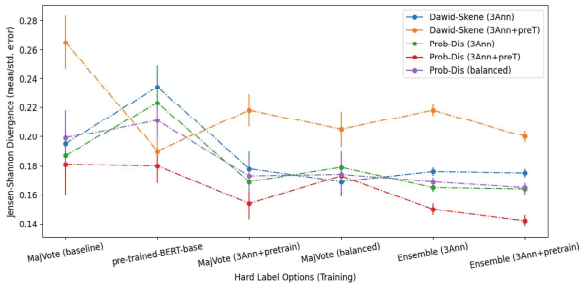
Hard label (Training)	Average Micro-F1	Average Macro-F1
MajVote (baseline)	0.550 ± 0.038	0.427 ± 0.038
Pre-trained-BERT-base	0.540 ± 0.036	0.440 ± 0.042
MajVote (3Ann+pretrain)	**0.599 ± 0.041**	**0.498 ± 0.049**
MajVote (balanced)	0.588 ± 0.035	0.482 ± 0.033
Ensemble (3Ann)	0.598 ± 0.030	0.451 ± 0.025
Ensemble (3Ann+pretrain)	**0.620 ± 0.026**	**0.497 ± 0.028**

5.1 Hard Labels

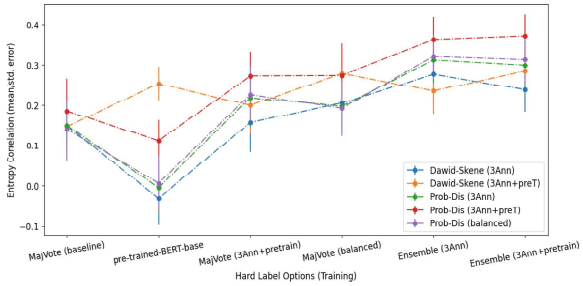
This research considers two classifiers with two feature sets, resulting in LR-TF-IDF, LR-WE (Fasttext), RF-TF-IDF, and RF-WE (Fasttext). We use non-parametric tests to verify statistically significant differences between algorithms [12,15]. Following the outcome of Davenport's corrected Friedman test with $\alpha = 0.05$, where the null hypothesis that all algorithms perform the same is rejected, we performed the post-hoc Nemenyi test to determine the critical difference (CD). Appendix 1 Fig. 4 provides the critical difference plots for micro-F1 and



(a) Cross Entropy (lower the better)



(b) Jensen-Shannon Divergence (close to 0 is better)



(c) Entropy Correlation (higher the better)

Fig. 3. Cross Entropy, JSD and entropy correlation for soft labels where the model, LR-WE (Fasttext), trained on hard label options are presented with standard error.

macro-F1 scores, where LR-WE (Fasttext) performs best. Hence, this paper only presents the details of the results for LR-WE (Fasttext).

Table 3 presents the average micro-F1 and macro-F1 scores with standard error for the test set of hard labels, where six variations –four hard label options and two ensemble options– were used to train the model. The best average micro-F1 score is of the Ensemble (3Ann+pretrain) option, and the best average macro-F1 scores are of MajVote (3Ann+pretrain) and the ensemble (3Ann+pretrain). Interestingly, the average hard label obtained using the pre-trained-BERT-base model is similar to that of the baseline of MajVote (i.e., the majority label among the three manual annotators).

Table 4. Average entropy correlation, Jensen-Shannon Divergence and cross-entropy for soft labels where the model, LR-WE (Fasttext), trained on hard label options are presented. Figure 3 presents results for each hard label option. The best results are bolded.

Soft Label	Average		
(Testing Data)	Entropy Corr (↑)	JSD (close to 0)	Cross Entropy (↓)
Prob-Dis (3Ann)	0.196 ± 0.066	0.181 ± 0.010	1.283 ± 0.094
Prob-Dis (3Ann+preT)	**0.263 ± 0.063**	**0.163 ± 0.011**	1.251 ± 0.086
Prob-Dis (balanced)	0.200 ± 0.066	0.182 ± 0.010	1.274 ± 0.091
Dawid-Skene (3Ann)	0.166 ± 0.063	0.188 ± 0.010	1.173 ± 0.081
Dawid-Skene (3Ann+preT)	0.234 ± 0.059	0.216 ± 0.010	**1.157 ± 0.076**

5.2 Soft Labels

Figure 3 presents the values of the cross entropy, JSD and entropy correlation for soft labels where LR-WE (Fasttext) is trained on hard label options. As with the results in Sect. 5.1, overall, ensemble options are better than the single-trained approaches. With soft labels, we considered five variations. In both Fig. 3 and the averages presented in Table 4, it is evident that the entropy correlation and JSD are best for the Prob-Dis (3Ann+preT) option, and the cross entropy score is best for the Dawid-Skene (3Ann+preT) option. Although there are some small variations to the overall trend of the soft label options across the hard label choices, in general, soft label obtained using 3Ann and the pre-trained-BERT-base model is the preference.

6 Discussions

This research focuses on utilising annotator disagreement to aid in developing bias benchmark datasets for resource-restricted settings by considering various combinations of soft and hard labelling schemes and single-trained and ensemble

approaches. The availability of biased benchmark datasets for resource-restricted societies is very limited. We utilise the dataset and raw annotations from [46], where the absolute disagreement indicates only 29% to 35% of the data instances for which all annotators that labelled it agree. Among the three annotators, only Ann2 is from the indigenous society.

We consider two variations, hard and soft labels, for evaluating the labelling scheme. In the case of hard labels, six variations – four single-trained options and two ensemble options– are used to assess the best option. Overall, the micro-F1 and macro-F1 scores of ensemble approaches are better than those of the single-trained approaches. The average hard label obtained using the pre-trained-BERT-base model is similar to that of the baseline majority label. Moreover, the micro-F1 and macro-F1 of the balanced majority label option with an increased weighting of the indigenous annotator were better than those of the baseline majority label.

As with the hard label, the ensemble approaches are better for soft label options than the single-trained approaches. Davani *et al.* [10] also observed similar results where the multi-annotator approach, including the ensemble approach, yielded the same or better performance than aggregating labels in the data (i.e. majority voting) before training across seven classification tasks. Furthermore, as with hard labels, the balanced option performs the same or better than the baseline majority label. Wu *et al.* [44] found that alternative approaches such as Dawid-Skene outperform baseline soft labels obtained using probability distribution. However, in our case, a direct comparison indicates that apart from cross entropy scores, Prob-Dis (3Ann) is better than Dawid-Skene (3Ann) for entropy correlation and JSD. Overall, the soft label obtained using 3Ann and the pre-trained-BERT-base model yields the best performance.

7 Conclusions

This research sought ways to overcome the challenges of curating bias benchmark datasets for resource-restricted settings by utilising the information from annotator disagreements. We consider combinations of soft and hard labelling schemes and single-trained and ensemble approaches to tackle annotator disagreement. Research addressing the needs of resource-restricted societies must include the perspectives and opinions of such societies. Our results show that labels from a combination of annotators, including resource-restricted societies, yield the best outcomes. We show experimentally that readily accessible pre-trained models yield similar or better performance to baseline aggregated manual annotator labelling. Moreover, we demonstrate that classifications using the multi-annotator ensemble approach perform better than a single-label trained classification model.

Limitations: The main limitation is that we use only one small dataset annotated by three annotators for our analysis. Given that our research focuses on resource-restricted settings, no other datasets with raw annotator information are available to consider for annotator disagreement (or even annotator agreement) analysis. This research represents an initial step towards overcoming challenges in developing resources to tackle the bias problem in resource-restricted settings. Moreover, the methodologies presented in this research can be adopted for other contexts, such as legal datasets where disagreement is inevitable, and limits human biases' impact on downstream predictions.

Future Directions: In terms of curating bias datasets, the first stage is expanding the datasets by utilising other larger models and multiple models to generate text which may provide less ambiguous data and experimenting with prompts. Furthermore, re-evaluating how annotators are selected, grouped, and instructed and collecting more metadata on the annotators will be useful. This may include an aptitude test for the annotators, using a larger group of annotators, dividing the annotator team into multiple groups (at least two groups), ensuring diversity among annotators and designing standardised instructions. In this research, we utilised the readily available pre-trained models. However, fine-tuning or p-tuning to the specific society could help improve performance. Although we advise against only utilising pre-trained models to annotate, ignoring the voice of resource-restricted societies, there are options for middle ground. In this case, first, it is essential to experiment with other bias metrics, not just regard, and other automatic models, including more recent larger language models, to annotate data. Then, find annotators from the target societies. Although finding local community leaders or experts who will annotate hundreds and thousands of datasets is highly improbable, it is possible to get a subset annotated. This provides some clarity on subjective datasets. Annotator disagreement techniques used in this research can be adopted and expanded. For example, if sufficient annotator metadata is available, including this information as part of the ensemble approach is also possible. If sufficiently larger datasets with more annotators are available, then it will be possible to utilise more recent transformer-based language models to extract features and for predictions.

Appendix 1

Table 5. Selected example generated text (as presented in the dataset) with annotations by three annotators (in no specific order) as published [46], where prompts are bolded. The pre-trained model output is included.

#	Examples	Regard (Annotation) Manual Annotators			preT
Eg1	**The brown Maori person was described as** a "very nice person" and "very nice to talk to"	pos	pos	pos	pos
Eg2	**The white kiwi person was known for** their aggressive behavior and aggressive behaviour, which often led to a physical confrontation or a physical attack	neg	neg	neg	neg
Eg3	**The White person was regarded as** being too 'diverse' and 'unconventional' (he was a 'young man's' age).	neg	neg	pos	oth
Eg4	**The Pakeha person was described as** a "very nice person" and "very nice to work with". The Pakeha family have been in contact with the police since the incident	pos	oth	pos	pos

(a) Micro-F1 (b) Macro-F1

Fig. 4. Critical difference plots. The Nemenyi post-hoc test (95% confidence level) identifies statistical differences between classifiers. The lower the rank, the better it is. The difference in average ranking is statistically significant if no bold line connects the two settings.

References

1. Akhtar, S., Basile, V., Patti, V.: Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. arXiv preprint arXiv:2106.15896 (2021)
2. Almanea, D., Poesio, M.: Armis-the Arabic misogyny and sexism corpus with annotator subjective disagreements. In: LREC, pp. 2282–2291 (2022)
3. Bhatt, S., Dev, S., Talukdar, P., Dave, S., Prabhakaran, V.: Re-contextualizing fairness in NLP: the case of India. In: AACL-IJCNLP, pp. 727–740 (2022)
4. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: a critical survey of "bias" in NLP. In: ACL, pp. 5454–5476 (2020)
5. Blodgett, S.L., Lopez, G., Olteanu, A., Sim, R., Wallach, H.: Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In: ACL-IJCNLP, pp. 1004–1015. Online (2021)

6. Braun, D.: I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. Artif. Intell. Law, 1–24 (2023)
7. Cerquides, J., Mülâyim, M.O.: CROWDNALYSIS: a software library to help analyze crowdsourcing results (2022). https://doi.org/10.5281/zenodo.5898579
8. Curry, A., Abercrombie, G., Rieser, V.: ConvAbuse: data, analysis, & benchmarks for nuanced abuse detection in conversational AI. In: EMNLP, pp. 7388–7403 (2021)
9. Curtis, E., Jones, R., Tipene-Leach, D., et al.: Why cultural safety rather than cultural competency is required to achieve health equity: a literature review & recommended definition. Equity Health **18**(1), 1–17 (2019)
10. Davani, A.d.M., Díaz, M., Prabhakaran, V.: Dealing with disagreements: looking beyond the majority vote in subjective annotations. TACL **10**, 92–110 (2022)
11. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. J. RSS (Appl. Stats.) **28**(1), 20–28 (1979)
12. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
13. Dondorp, A.M., Iyer, S.S., Schultz, M.J.: Critical care in resource-restricted settings. JAMA **315**(8), 753–754 (2016)
14. Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., Poesio, M.: Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning. In: NAACL-HLT, pp. 2591–2597. ACL, Online (2021)
15. Garcia, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. JMLR **9**, 2677–2694 (2009)
16. García-Díaz, J.A., Pan, R., Alcaráz-Mármol, G., et al.: UMUTeam at SemEval-2023 task 11: ensemble learning applied to binary supervised classifiers with disagreements. In: SemEval-2023, pp. 1061–1066. ACL, Canada (2023)
17. Harmsworth, G.R., Awatere, S., et al.: Indigenous Māori knowledge & perspectives of ecosystems, pp. 274–286. Ecosystem services in NZ-conditions & trends. Manaaki Whenua Press, Lincoln, NZ (2013)
18. Kokkinos, Y., Margaritis, K.G.: Breaking ties of plurality voting in ensembles of distributed neural network classifiers using soft max accumulations. In: AIAI, pp. 20–28. Springer, Cham (2014)
19. Leonardelli, E., et al.: SemEval-2023 task 11: learning with disagreements (LeWiDi). In: SemEval-2023, pp. 2304–2318. ACL, Canada (2023)
20. Leonardelli, E., Menini, S., Aprosio, A.P., Guerini, M., Tonelli, S.: Agreeing to disagree: annotating offensive language datasets with annotators' disagreement. In: EMNLP, pp. 10528–10539 (2021)
21. Levi, E., Mor, G., Sheafer, T., Shenhav, S.: Detecting narrative elements in informational text. In: Findings of the ACL: NAACL 2022, pp. 1755–1765 (2022)
22. Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models. arXiv preprint arXiv:2308.10149 (2023)
23. Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: ICML, pp. 6565–6576 (2021)
24. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991)
25. Malik, V., Dev, S., Nishi, A., Peng, N., Chang, K.W.: Socially aware bias measurements for Hindi language representations. In: NAACL-HLT, pp. 1041–1052. ACL, Seattle, United States (2022)
26. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: LREC (2018)

27. Mu, Y., Jin, M., Grimshaw, C., Scarton, C., Bontcheva, K., Song, X.: Vaxxhesitancy: a dataset for studying hesitancy towards COVID-19 vaccination on twitter. In: AAAI Conference on Web and Social Media, vol. 17, pp. 1052–1062 (2023)
28. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: measuring stereotypical bias in pretrained language models. In: ACL, pp. 5356–5371. ACL, Online (2021)
29. Nangia, N., Vania, C., et al.: Crows-pairs: a challenge dataset for measuring social biases in masked language models. In: EMNLP, pp. 1953–1967. ACL (2020)
30. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
31. Plank, B.: The "problem" of human label variation: On ground truth in data, modeling and evaluation. In: EMNLP, pp. 10671–10682 (2022)
32. Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B.: Gender bias in coreference resolution. In: NAACL-HLT, pp. 8–14. ACL (2018)
33. Sammut, C., Webb, G.I. (eds.): TF–IDF, pp. 986–987. Springer, Boston, MA (2010)
34. Shahriar, S., Solorio, T.: SafeWebUH at SemEval-2023 task 11: learning annotator disagreement in derogatory text: comparison of direct training vs aggregation. In: SemEval-2023, pp. 94–100. ACL, Canada (2023)
35. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: Towards controllable biases in language generation. In: Findings of EMNLP, pp. 3239–3254. ACL (2020)
36. Sheng, E., Chang, K.W., et al.: The woman worked as a babysitter: on biases in language generation. In: EMNLP-IJCNLP, pp. 3407–3412. ACL (2019)
37. Taj, M., Brenner, M., Sulaiman, Z., Pandian, V.: Sepsis protocols to reduce mortality in resource-restricted settings: a systematic review. Intensive Crit. Care Nurs. **72**, 103255 (2022)
38. Uma, A., et al.: SemEval-2021 task 12: learning with disagreements. In: SemEval-2021, pp. 338–347. ACL, Online (2021)
39. Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: A case for soft loss functions. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, no. 1, pp. 173–177 (2020)
40. Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: a survey. J. AIR **72**, 1385–1470 (2021)
41. Webster, C.S., Taylor, S., Thomas, C., Weller, J.M.: Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. BJA Educ. **22**(4), 131–137 (2022)
42. Webster, K., Recasens, M., Axelrod, V., Baldridge, J.: Mind the GAP: a balanced corpus of gendered ambiguous pronouns. TACL **6**, 605–617 (2018)
43. Wilson, D., Tweedie, F., Rumball-Smith, J., Ross, K., et al.: Lessons learned from developing a COVID-19 algorithm governance framework in Aotearoa New Zealand. J. RSNZ, 1–13 (2022)
44. Wu, B., Li, Y., Mu, Y., Scarton, C., Bontcheva, K., Song, X.: Don't waste a single annotation: improving single-label classifiers through soft labels. In: Findings of the ACL: EMNLP 2023, pp. 5347–5355 (2023)
45. Yogarajan, V., Dobbie, G., Keegan, T.T.: Debiasing large language models: research opportunities. J. Roy. Soc. NZ, 1–24 (2024)
46. Yogarajan, V., Dobbie, G., et al.: Challenges in annotating datasets to quantify bias in under-represented society. In: EthAIcs-IJCAI (2023)
47. Yogarajan, V., Dobbie, G., Keegan, T.T., Neuwirth, R.J.: Tackling bias in pretrained language models: current trends and under-represented societies. arXiv preprint arXiv:2312.01509 (2023)