Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations

Aida Mostafazadeh Davani
University of Southern California, USA
mostafaz@usc.edu

Mark Díaz
Google Research, USA
markdiaz@google.com

Vinodkumar Prabhakaran
Google Research, USA
vinodkpg@google.com

Abstract

Majority voting and averaging are common approaches used to resolve annotator disagreements and derive single ground truth labels from multiple annotations. However, annotators may systematically disagree with one another, often reflecting their individual biases and values, especially in the case of subjective tasks such as detecting affect, aggression, and hate speech. Annotator disagreements may capture important nuances in such tasks that are often ignored while aggregating annotations to a single ground truth. In order to address this, we investigate the efficacy of multi-annotator models. In particular, our multi-task based approach treats predicting each annotators' judgements as separate subtasks, while sharing a common learned representation of the task. We show that this approach yields same or better performance than aggregating labels in the data prior to training across seven different binary classification tasks. Our approach also provides a way to estimate uncertainty in predictions, which we demonstrate better correlate with annotation disagreements than traditional methods. Being able to model uncertainty is especially useful in deployment scenarios where knowing when not to make a prediction is important.

1 Introduction

Obtaining multiple annotator judgements on the same data instances is a common practice in NLP in order to improve the quality of final labels (Snow et al., 2008; Nowak and Rüger, 2010). In case of disagreements between annotations, they are often aggregated by majority voting, averaging (Sabou et al., 2014), or adjudicating by an ''expert'' (Waseem and Hovy, 2016), to derive a singular ground truth or gold label that is later used for training supervised machine learning models. However, in many subjective tasks, there often exists no such single ''right'' answer (Alm, 2011) and enforcing a single ground truth sacrifices the valuable nuances embedded in annotator's assessments of the stimuli and their disagreements (Aroyo and Welty, 2013; Cheplygina and Pluim, 2018).

Annotators' sociodemographic factors, moral values, and lived experiences often influence their interpretations, especially in subjective tasks such as identifying political stances (Luo et al., 2020), sentiment (Díaz et al., 2018), and online abuse and hate speech (Cowan and Khatchadourian, 2003; Waseem, 2016; Patton et al., 2019). For instance, Waseem (2016) found that feminist and anti-racist activists systematically disagree with crowd workers on their hate speech annotations. Similarly, annotators' political affiliation affects how they annotate the neutrality of political stances (Luo et al., 2020). An adverse effect of majority vote in such cases is limiting representation of minority perspectives in data (Prabhakaran et al., 2021), potentially reinforcing societal disparities and harms.

Another consequence of majority voting, when applied to subjective annotations, is that the resulting labels may not be internally consistent. For example, consider a scenario where a sentence in a hate-speech dataset is annotated by a set of annotators, the majority of whom consider a phrase in it to be offensive, yet another sentence with the same phrase is annotated by a different set of annotators, the majority of whom *do not* find the phrase to be offensive. Upon majority vote, the first sentence would be labeled as hate speech and the second sentence would not, despite containing similar content. Such inconsistencies in the majority label will add noise to the learning step, while the systematicity in the individual annotations is lost.

Finally, majority vote and similar aggregation approaches assume that an annotator's judgements about different instances are independent from one

Transactions of the Association for Computational Linguistics, vol. 10, pp. 92–110, 2022. https://doi.org/10.1162/tacl_a_00449
Action Editor: Dirk Dovy. Submission batch: 5/2021; Revision batch: 8/2021; Published 1/2022.

another. However, as outlined above, annotators' decisions are often correlated, reflecting their subjective biases. Prior work has investigated Bayesian methods to account for such systematic differences between annotators (Paun et al., 2018), however, they approach this as an alternate means to derive a single ground truth label, thereby masking the degree to which annotators disagreed.

Our proposed solution is simple: We introduce multi-annotator architectures to preserve and model the internal consistency in each annotators' labels as well as their systematic disagreements with other annotators. We show that the multi-task framework (Liu et al., 2019) provides an efficient way to implement a multi-annotator architecture that captures the differences between individual annotators' perspectives using the subset of data instances they labeled, while also benefiting from the shared underlying layers fine-tuned for the task using the entire dataset. Preserving different annotators' perspectives until the prediction step provides better flexibility for downstream applications. In particular, we demonstrate that it provides better estimates for uncertainty in predictions. This will improve decision making in practice—for instance, to determine when not to make a prediction or when to recommend a manual review.

Our contributions in this paper are three-fold: (1) We develop an efficient multi-annotator strategy that matches or outperforms baseline models on seven different subjective tasks by preserving annotators' individual and collective perspectives throughout the training process. (2) We obtain an interpretable way to estimate model uncertainty that better correlates with annotator disagreements than traditional uncertainty estimates across all seven tasks. (3) We demonstrate that model uncertainty correlates with certain types of error, providing a useful signal to avoid erroneous predictions in real-world deployments.

2 Literature Review

Learning to recognize and interpret subjective language has a long history in NLP (Wiebe et al., 2004; Alm, 2011). While all human judgements embed some degree of subjectivity, it is commonly agreed that certain NLP tasks tend to be more subjective in nature. Examples of such relatively subjective tasks include sentiment analysis (Pang and Lee, 2004; Liu et al., 2010), affect modeling (Alm, 2008; Liu et al., 2003), emotion detection (Hirschberg et al., 2003; Mihalcea and Liu, 2006), and hate speech detection (Warner and Hirschberg, 2012). Alm (2011) argues that achieving a single *real ''ground truth''* is not possible, nor essential, in subjective tasks, and calls for finding ways to model subjective interpretations of annotators, rather than seeking to reduce the variability in annotations. While which NLP tasks count as subjective may be contested, we focus on two tasks that are markedly subjective in nature.

2.1 Detecting Online Abuse

NLP-aided approaches to detect abusive behavior online is an active research area (Schmidt and Wiegand, 2017; Mishra et al., 2019; Corazza et al., 2020). Researchers have developed typologies of online abuse (Waseem et al., 2017), constructed datasets annotated with different types of abusive language (Warner and Hirschberg, 2012; Price et al., 2020; Vidgen et al., 2021), and built NLP models to detect them efficiently (Davidson et al., 2017; Mozafari et al., 2019). Researchers have also expanded the focus to more subtle forms of abuse such as condescension and microaggressions (Breitfeller et al., 2019; Jurgens et al., 2019).

However, recent research has demonstrated that these models tend to reflect and propagate various societal biases, causing disparate harms to marginalized groups. For instance, toxicity prediction models were shown to have biases towards mentions of certain identity terms (Dixon et al., 2018), specific named entities (Prabhakaran et al., 2019), and disabilities (Hutchinson et al., 2020). Similarly these models are shown to overestimate the prevalence of toxicity in African American Vernacular English (Sap et al., 2019; Davidson et al., 2019; Zhou et al., 2021). Most of these studies demonstrate association biases present in data; for instance, Hutchinson et al. (2020) show that discussions about mental illness are often associated with topics such as gun violence, homelessness, and drugs, likely the reason for the learned association of mental illness related terms with toxicity. While whether a piece of text is hateful or not depends also on the context (Prabhakaran et al., 2020), not much work has investigated the human annotator biases present in the training labels, and how they impact downstream predictions.

2.2 Detecting Emotions

Detecting emotions from language has been a significant area of research in NLP for the past two decades (Liscombe et al., 2003; Aman and Szpakowicz, 2007; Desmet and Hoste, 2013; Hirschberg and Manning, 2015; Poria et al., 2019). Annotated datasets used for training emotion detection models vary across domains, and use different taxonomies of emotions. While several datasets (Strapparava and Mihalcea, 2007; Buechel and Hahn, 2017) include a small set of labels representing the six Ekman emotions (Ekman, 1992—*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*), or bipolar dimensions of affect (*arousal* and *valence*; Russell, 2003), others such as Demszky et al. (2020) and Crowdflower (2016) include a wider range of emotion labels according to the Plutchik emotion wheel (Plutchik, 1980) or the complex semantic space of emotions (Cowen et al., 2019). Perceiving emotions is a subjective task affected by various contextual factors, such as time, speaker, mood, personality, and culture (Mower et al., 2009). Since aggregating annotations of emotion expressions loses such contextual nuances, some researchers provide a distributional representation of emotions (Fayek et al., 2016; Ando et al., 2018). Here, we use annotations for the six Ekman emotions present in the dataset released by Demszky et al. (2020) to demonstrate how our multi-annotator approach can capture emotions in a disaggregated fashion.

2.3 Annotation Disagreement

Researchers have studied different sources of annotator disagreements. Krippendorff (2011) argued that there are at least two types of disagreement in content coding: random variation, which comes as an unavoidable by-product of human coding, and systematic disagreement, which is influenced by features of the data or annotators. Dumitrache (2015) identifies different sources of disagreement as (a) the clarity of an annotation label (i.e., task descriptions), (b) the ambiguity of the text, and (c) differences in workers. Aroyo and Welty (2013) also studied inter-annotator disagreement in association with features of the input, showing that it reflects semantic ambiguity of the training instances. Textual features have been shown to predict annotators' disagreement in determining the meaning of ambiguous words (Alonso et al., 2015). Acknowledging inter-annotator disagreement as an indicator of annotator differences, Kairam and Heer (2016) clustered crowd-workers based on their annotation behaviors, and proposed a method for interpreting annotation disagreements and its sources.

For highly subjective tasks such as hate speech and emotions detection, annotation disagreements can be rooted in the differing subjectivities and value systems of annotators. In these cases, annotators build a subjective social reality as a basis for social judgements and behaviors (Greifeneder et al., 2017), which explains their labeling procedure. For example, in interviews with annotators in an aggression labeling task, Patton et al. (2019) found that expert annotators from communities discussed in gang-related tweets drew on their lived experience to produce different label judgements compared with graduate student researchers. Such annotators whose lived experiences bring important perspectives to the task would be dramatically underrepresented on generic crowd work platforms and, by definition, would be outvoted in disagreements subject to majority vote. Majority vote also necessarily obfuscates differences among groups underrepresented in annotator pools, such as older adults who can exhibit views on aging distinct from crowd workers (Díaz, 2020), the majority of whom tend to be younger (Ross et al., 2010).

Some studies have proposed alternatives to majority voting when aggregating multiple annotations. In early work, Dawid and Skene (1979) used the *EM* algorithm to obtain maximum likelihood estimates of the ''true'' label to account for annotator errors. De Marneffe et al. (2012) used the individual annotation distributions to predict areas of uncertainty in veridicality assessment. Hovy et al. (2013) proposed an approach based on item-response model that uses posterior entropy to choose which annotators are trustworthy. Waterhouse (2013) developed a pointwise mutual information metric to quantify the amount of information in an annotator's judgement that can be used to estimate the ''correct'' label of an instance. Gordon et al. (2021) explore multiple annotators judgements to disentangle stable opinions from noise by estimating intra-annotator consistency. All these approaches aim to obtain the ''correct'' label, accounting for erroneous or non-trustworthy annotators, whereas we focus on retaining the annotator disagreements through the modeling process.

A few studies have explored approaches for utilizing annotation disagreement during model training. Prabhakaran et al. (2012) explored applying higher cost for errors made on unanimous annotations to decrease the penalty of mis-labeling inputs with higher disagreement. Similarly, (Plank et al., 2014) incorporated annotator disagreement into the loss function of a structured perceptron model for better predicting part-of-speech tags. Our work also utilizes annotator disagreements rather than resolving them in the data stage; however, we use a multi-task architecture using a shared representation to model annotator disagreements, rather than using it in loss function. Cohn and Specia (2013) use a multi-task approach to model annotator differences in machine translation annotations. While they use a Gaussian Process approach, we use the multi-task approach on top of pre-trained language models (Liu et al., 2019). Chou and Lee (2019) proposed an approach where they model individual annotators separately in an inner layer to improve the final prediction. In contrast, our method uses the multi-task architecture, and provides the additional ability to utilize multiple predictions during deployment, for instance, to measure uncertainty. Fornaciari et al. (2021) also leveraged annotator disagreement using a multi-task model that adds an auxiliary task to predict the soft label distribution over annotator labels, which improves the performance even in less subjective tasks such as part-of-speech tagging. In contrast, our approach models several annotators' labels as multiple tasks and obtains their disagreement.

2.4 Prediction Uncertainty

Model uncertainty denotes the confidence of model predictions, which has specific applications in non-deterministic machine learning tasks. For instance, interpreting model outputs and its confidence is critical in autonomous vehicle driving, where wrong predictions are costly or harmful (Schwab and Karlen, 2019). In subjective tasks, uncertainty embeds additional information that supports result interpretation (Ghandeharioun et al., 2019). For example, the level of uncertainty could help determine when and how moderators take part in a human-in-the-loop content moderation (Chandrasekharan et al., 2019; Liu, 2020).

The simplest approach for uncertainty estimation is through prediction probability from a Softmax distribution (Hendrycks and Gimpel, 2017). However, as the input data gets farther from the training data, this probability estimation naturally yields extrapolations with unsupported high confidence (Gal and Ghahramani, 2016). Instead, Gal and Ghahramani (2016) proposed the *Monte Carlo* dropout approach to estimate uncertainty by iteratively applying dropouts to all layers of the model and calculating the variance of generated outputs. Such estimations based on the probability of a single ground truth label overlooks the many factors that contribute to uncertainty (Kläs and Vollmer, 2018). In contrast, Passonneau and Carpenter (2014) demonstrate the benefits of measuring uncertainty for the ground truth label by fitting a probabilistic model to individual annotators' observed labels. Similarly, we demonstrate that calculating annotation disagreement by predicting a set of annotations for the input yields a better estimation of uncertainty than estimations based on the probability of the majority label.

3 Methodology

We define the classification task on an annotated dataset $D = (X, A, Y)$, in which X is a set of text instances, A is the set of annotators and Y is the annotation matrix, in which each entry $y_{ij} \in \{0, 1\}$ represents the label assigned to $x_i \in X$ by $a_j \in A$. In most annotated datasets Y includes many missing values, because each annotator only labels a subset of all instances. We use $\bar{y_{i,}}$ to refer to the annotations present for item x_i. Similarly, we use $\bar{y_{,j}}$ to refer to the annotations made by annotator a_j. The classification task aims to predict $maj(\bar{y_{i,}}) \in \{0, 1\}$, which is the label assigned to x_i based on the majority vote over $\bar{y_{i,}}$. We use majority vote, the most commonly used aggregation method; however, our proposed approach leaves open the choice of the aggregation method depending on deployment contexts.

We consider three different multi-annotator architectures: *ensemble*, *multi-label*, and *multi-task*. Figure 1 shows the schematic differences between these three variations. All variations use Bidirectional Encoder Representations from Transformers (BERT-base; Devlin et al., 2019). For each instance x_i, a generic representation $h_i \in \mathbb{R}^d$ is generated by the pre-trained BERT-base, and then fine-tuned along with other

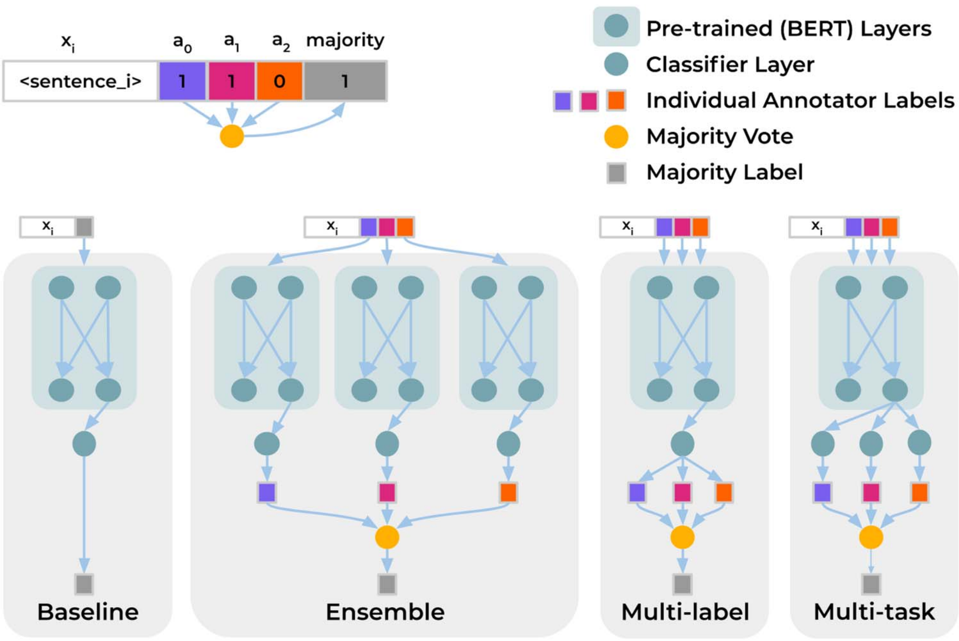

Figure 1: Comparison between approaches for multi-annotator model (ensemble, multi-label, and multi-task) and majority label prediction (baseline). Annotation prediction models are trained based on all annotations and apply majority voting to predict the final label.

components of the classifier during training. The size of the representation vector, d, is defined by the BERT configuration and is set to 768 for the pre-trained BERT-base. While our experiments are all performed with BERT-base, our methods are not restricted to BERT in their nature, and can be implemented with other pre-trained language models, for example, RoBERTa (Zhu et al., 2020).

3.1 Baseline Model Using Majority Labels

The baseline model captures the most common approach: a single-task classifier trained to predict the aggregated label for each instance (i.e., majority vote, in our case). It is built by adding a fully connected layer to BERT-base outputs (h_i). The fully connected layer applies a linear transformation followed by Softmax function to generate the probability of the majority label, $P(maj(\bar{y_{i,}})|h_i)$. Compared to the other models described in this section, the baseline model does not make use of annotation matrix Y, as it directly predicts the aggregated label $maj(\bar{y_{i,}})$.

3.2 Ensemble Approach

An intuitive approach towards multi-annotator models might be to train an ensemble of models, each trained on different annotators' labels. This approach is not always practical, as it may increase the training time prohibitively. The ensemble approach applies $|A|$ single-task classifiers, each for training and predicting the annotations generated by one annotator. During training, the j-th classifier is independently fine-tuned to predict $\bar{y_{,j}}$, which includes all annotations provided by the j-th annotator. During test time, we aggregate the outputs by the majority vote of all $|A|$ models to predict $P(maj(\bar{y_{i,}})|x_i)$.[1]

3.3 Multi-label Approach

A more practical approach for multi-annotator modeling is to consider the problem as a multi-label problem where each label denotes individual annotators' labels. More specifically, the multi-label approach attempts to learn to predict $|A|$ labels for each input using a multi-label classification framework. The model first adds a fully connected layer to transform each h_i to a $|A|$-dimensional vector, and then applies a Sigmoid function to the j-th dimension to generate y_{ij}. Since Y includes many missing values, the classification loss is calculated based on the available labels $y_{ij} \in \bar{y_{i,}}$. However, during test time, all $|A|$ outputs are aggregated to predict $P(maj(\bar{y_{i,}})|x_i)$.

[1] During prediction, multi-annotator models do not have access to the list of annotators who originally provided the labels for each instance. Therefore, the original majority vote is predicted as the majority vote among all annotators.

3.4 Multi-task Approach

The multi-task based multi-annotator approach attempts to learn multiple annotators' perspectives (labels) as separate classification tasks, all of which share encoder layers to generate the same representation of the input sentence h_i, each with its separate fully connected layer and softmax activation. Compared with the multi-label approach, the multi-task model includes a fully connected layer explicitly fine-tuned for each annotator. However, compared with the ensemble approach, the representation layers which generate h_i are fine-tuned based on the outputs of all annotation tasks. The loss function is created as the summation of all available labels $\bar{y_{i,}}$ for each instance x_i. During test time, the model considers the outputs of all annotation tasks to predict the majority label $P(maj(\bar{y_{i,}})|x_i)$.

4 Experiments

4.1 Data

For this study, we perform experiments on two datasets annotated for subjective tasks: Gab Hate Corpus (GHC; Kennedy et al., 2020) and GoEmotions dataset (Demszky et al., 2020). Both datasets capture per-annotator labels for instances along with corresponding annotators' anonymous ID, allowing us to model each annotator separately.

4.1.1 Gab Hate Corpus (GHC)

GHC (Kennedy et al., 2020), includes $|X| =$ 27,665 social-media posts collected from a public corpus of Gab.com (Gaffney, 2018), each annotated for whether or not they contain hate speech. Kennedy et al. (2020) define hate speech as language that dehumanizes, attacks human dignity, derogates, incites violence, or supports hateful ideology, such as white supremacy. Each instance in GHC is annotated by at least three annotators from a set of $|A| = 18$ annotators. The number of annotations varies for each instance ($M(|\bar{y_{i,}}|) = 3.13$, $SD(|\bar{y_{i,}}|) = 0.39$), and in total, there are 86,529 annotations. The number of annotated instances per annotator also varies significantly ($M(\bar{y_{,j}}) = 4807.17$, $SD(\bar{y_{,j}}) = 3184.89$).

4.1.2 GoEmotions

We use a subset of the GoEmotions dataset (Demszky et al., 2020) which contains Reddit posts annotated for 28 emotions, split across predefined train ($|X|_{train}$ = 43,410), test ($|X|_{test}$ = 5,427), and validation ($|X|_{val}$ = 5,426) subsets. Our experiments focus on the emotion annotations for the six Ekman (Ekman, 1992) emotions:*anger*, *disgust*,*fear*, *joy*, *sadness*, and *surprise*. Each instance in GoEmotions is annotated by three to five annotators from a set of $|A|$ = 82 annotators. The number of annotations varies for each instance ($M(|\bar{y_{i,}}|) = 3.58$, $SD(|\bar{y_{i,}}|) = 0.91$), and in total, there are 194,412 annotations. The number of annotated instances varies significantly across annotators ($M(\bar{y_{,j}}) = 2370.88$, $SD(\bar{y_{,j}}) = 2180.02$).

4.2 Experimental Setup

We implemented the classification models using the `transformers` (v3.1) library from HuggingFace (Wolf et al., 2020). The training steps employ the *Adam* optimizer (Kingma and Ba, 2015). Our experiment settings are configured similar to Kennedy et al. (2020) and Demszky et al. (2020), GHC experiments are conducted with a learning rates of $e-7$ and are trained for three epochs, whereas experiments on GoEmotions apply early stopping with a learning rate of $5e-6$. Since GHC does not have specific train and test subsets, we conducted 5 iterations of stratified 5-fold cross-validations for evaluation, changing only the random state for each iteration. GoEmotions experiments are performed as six different binary classification tasks, also repeated for 5 iterations, using the pre-defined train and test sets.

4.3 Results on GHC

4.3.1 Prediction Results

Table 1 reports the average and standard deviation of the precision, recall, and F$_1$-scores for various models, across the 5 iterations. The baseline model, which is trained using the majority vote as ground truth, is also tested against the majority vote labels. For the ensemble, multi-label, and multi-task models, we conduct two types of evaluation: First, we test how well the majority vote of predicted labels match the majority vote of annotations (columns 2-4 in Table 1); second, we report how well the individual predicted labels

	Majority Vote			Individual Labels		
Model	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	49.53±3.8	**68.78**±4.4	57.32±1.2	–	–	–
Ensemble	63.98±1.1	46.09±1.9	53.54±1.0	60.92±0.7	60.97±0.8	60.94±0.3
Multi-label	**66.02**±2.2	50.16±2.0	56.94±1.0	**67.22**±1.4	55.33±2.0	60.65±0.7
Multi-task	59.03±0.9	59.98±0.6	**59.49**±0.2	63.71±1.3	**62.76**±1.5	**63.20**±0.3

Table 1: The average and standard deviation of precision, recall, and F-score of model predictions on the **GHC** dataset, evaluated during 5 iterations of 5-fold stratified cross validation. Majority Vote section represent models' performance on predicting the majority vote, while Individual Labels section reports performance on predicting each raw annotation.

for each instance match the annotations (where available) by annotators (columns 5-7 in Table 1).

We observe that the ensemble model performs significantly worse ($\mathrm{F}_1 = 53.54$) than the baseline single-task model ($\mathrm{F}_1 = 57.32$) in predicting majority label. This is presumably due to the fact that each base model in the ensemble is trained using only the examples labeled by the corresponding annotator. Since the number of annotations varies significantly for different annotators (see Section 4.1), many base models end up with lower performance, resulting in lower overall performance.

Multi-label and multi-task models share most layers across different annotator heads. Thus, each annotator head benefits from the updates to the shared layers owing to all instances, regardless of whether they annotated it or not. The multi-label model performs slightly worse ($\mathrm{F}_1 = 56.94$) than the baseline model. In contrast, the multi-task model, which has a fully connected layer fine-tuned for each annotator, posted a significantly higher F-score ($\mathrm{F}_1 = 59.49$) than the baseline model. In other words, fine-tuning each annotator head separately and then taking the majority vote performs better than taking the majority vote first and then training on that noisier label.

Moreover, the baseline model yields higher performance variance among different iterations, such that its standard deviations of precision, recall, and F_1 exceeds those of the other three methods. One possible explanation is that aggregating annotations based on majority votes disposes of information about each annotator and inserts noise into the labels. In other words, modeling each annotator, and their presumable internal consistency, could lead to more stable prediction results. However, this hypothesis requires further investigation.

We now evaluate the individual predictions made by the multi-annotator model (prior to majority vote) on how well they match individual annotators' labels (Table 1). All three multi-annotator approaches obtain higher F_1-scores than how the baseline model does in predicting majority labels (note that these are different tasks, and not directly comparable). The multi-task model achieved the highest F_1-score of 63.20. The result suggests that the multi-task model benefits from fine-tuning annotators separately (thereby avoiding inconsistencies due to majority votes) as well as learning from all instances in a shared fashion.

4.3.2 Modeling Uncertainty

Next, we study how well we can model uncertainty in predictions. We compare uncertainty in predictions with annotator disagreement, measured as the variance of the annotations.

$$\sigma^2(\bar{y_{i,}}) = \frac{\sum[y_{ij} = 1] \sum[y_{ij} = 0]}{|\bar{y_{i,}}|^2} \tag{1}$$

Since the ensemble, multi-label, and multi-task models all make separate predictions corresponding to each annotator, we can calculate the uncertainty in predictions to be the variance of the predicted annotations for each instance x_i. However, modeling prediction uncertainty in the case of single predictions is an open question. We compare our results with other common approaches for estimating uncertainty in single-task predictions such as *Softmax* probability of the final output for predicting majority vote (Hendrycks and Gimpel, 2017), and Monte Carlo dropouts (Gal and Ghahramani, 2016), or *MC dropout*, which

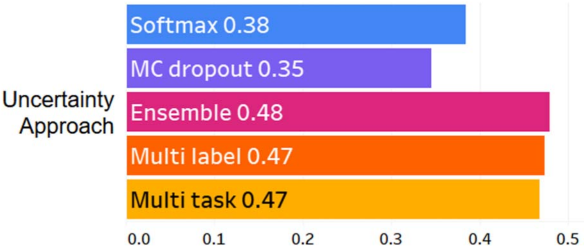

Figure 2: Correlation of different approaches for estimating prediction uncertainty with annotation disagreement on the **GHC**. Annotation modeling approaches better correlate with disagreement.

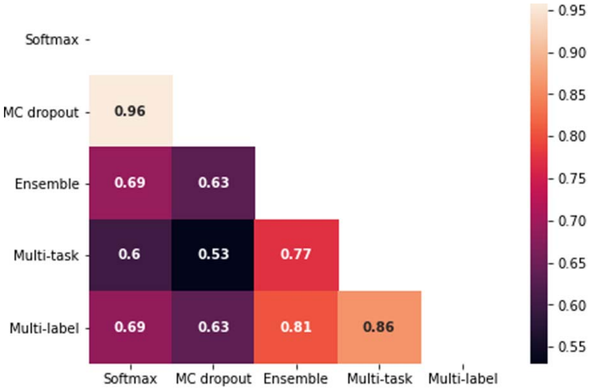

Figure 3: Correlation matrix of approaches for estimating uncertainty. MC dropout and Softmax have high correlation. Our multi-annotator models also have higher internal correlations.

iteratively applies dropouts to all layers of the model and calculates the variance in predictions.

Figure 2 shows the correlations of uncertainty estimation using each method with the annotation disagreement calculated as $\sigma^2(\bar{y_{i,}})$. While traditional estimations such as Softmax and MC dropout have a moderate correlation with annotator disagreements, the uncertainty measured by our three multi-annotator methods show significantly better correlation, with the ensemble method posting a slightly higher correlation than the other two methods. In other words, in addition to performing better on predicting majority votes, multi-annotator models also predict model uncertainty better than traditional approaches.

We further analyze the pair-wise correlation between estimations of uncertainty by different approaches (Figure 3). As expected, the Softmax and MC dropout methods are highly correlated, and similarly, our methods show high correlation among themselves. It is also interesting to note that the uncertainty estimated by our methods also correlate significantly with traditional methods (i.e., between 0.6 and 0.7), except for the multi-task method and MC Dropout method, which have a lower correlation of 0.53.

Models	Training Time (in mins)
Baseline	20.5
Ensemble	158.4
Multi-label	22.8
Multi-task	22.3

Table 2: Training time (in minutes); the time it takes to train each model on 80% of the GHC.

The fact that the uncertainty scores for multi-task and multi-label models are highly correlated with each other (0.86) suggests that they both identify textual features that cause disagreement. We verified this by training a separate model using the same BERT-based setup using Sigmoid activation to directly predict the annotator disagreement. The predicted uncertainty by this model obtained similar correlation with the annotator uncertainty (0.47) as the multi-task and multi-label models.

4.3.3 Computation Time

We now assess the computation cost associated with the different approaches. Table 2 shows the time it took to train a single cross-validation fold (i.e., 80% of the dataset). As expected, the ensemble approach takes the longest to train, as it require training $|A|$ different models (each with varying training set sizes), and the baseline takes the shortest time. Impressively, multi-label and multi-task models do not take significantly more time to train. In other words, while the multi-task model trains additional layers for annotators, it adds only a marginal computation cost to the baseline model.

4.4 Results on GoEmotions

In this section, we describe results obtained on the six binary classification tasks performed using the GoEmotions dataset. Since the multi-task approach obtained better performance overall on GHC, we report the results on only the multi-task approach here. We start by assessing how well the multi-annotator model matches the single-task performance of predicting the majority label. Table 3 reports the average and standard deviation of F_1-scores over 5 iterations of training and

	Full Dataset ($\|A\|$ = 82)		Subset ($\|A\|$ = 53)	
Emotion	Baseline	Multi-task	Baseline	Multi-task
Anger	**40.38**±4.4	39.01±6.4	41.95±6.1	**42.75**±4.4
Disgust	**38.79**±3.9	38.31±1.9	**37.72**±2.0	35.77±2.0
Fear	**58.96**±5.0	54.97±6.1	57.68±3.7	**58.58**±2.3
Joy	47.80±2.2	**49.53**±3.6	**47.45**±3.1	46.26±1.2
Sadness	49.22±5.2	**50.36**±3.2	47.55±5.4	**48.00**±3.4
Surprise	**40.96**±2.9	38.97±3.6	39.44±5.7	**40.22**±2.2

Table 3: The average and standard deviation of model prediction f-score on the **GoEmotions** dataset, evaluated across 5 iterations using the pre-defined train-test splits in the dataset.

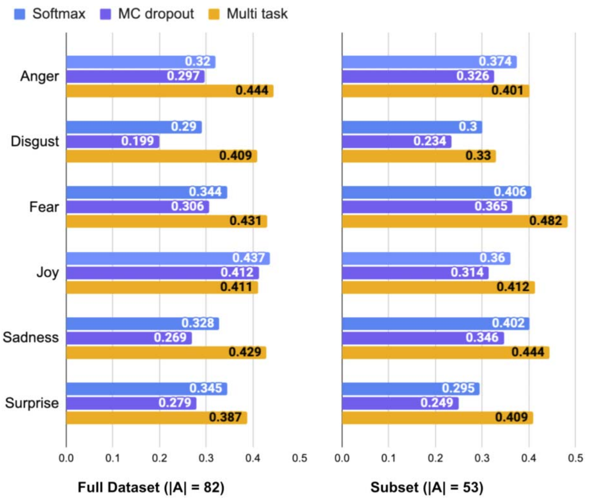

Figure 4: Correlation of different approaches for estimating prediction uncertainty with annotation disagreement for the **GoEmotions** dataset.

testing. Unlike GHC where we used 5-fold cross validation, for the GoEmotions dataset we use the pre-defined train, validation, test splits in the dataset. We verified that these splits are stratifed with respect to annotators. As in GHC experiments, while the baseline model is trained and tested on the majority vote, the multi-task model is trained on available annotator-level annotations for each instance and the predictions from all classifier heads are aggregated to get the final label during testing.

Results obtained on the full dataset are shown in the second and third columns of Table 3. While the multi-task model outperformed the baseline in predicting two emotions—*joy* and *sadness*—it underperformed the baseline for the other four emotions, although the ranges of F_1-scores largely overlap. It is also observed that the standard deviations of the multi-task model F_1-scores are significantly larger than what was observed for GHC.

On further inspection, we found that many annotators contributed very few annotations in the dataset. For instance, 29 annotators had fewer than 1000 annotations in the training set, six of them having fewer than 100. In addition, the label distribution is extremely skewed for all six emotions—ranging from 1.6% positive labels for *fear* on average across all annotators, to 4.0% positive labels on average for *joy*. Consequently, many annotator heads have too few positive instances to learn from; some had zero positive instances in the training set. This makes the corresponding learning tasks in the multi-task setting hard or even impossible on this dataset, and might explain the lower performance and higher variance in F_1-scores.

In order to make a fairer comparison, we performed our experiments on a subset of the dataset that only includes the annotations by 53 annotators who had more than 1000 annotations. Results obtained on this subset are in the fourth and fifth columns of Table 3. Our multi-annotator model outperforms the baseline model on predicting the majority label in four of the six tasks (*anger*, *fear*, *sadness*, and *surprise*), while obtaining slightly lower results on *disgust* and *joy*. While F_1-score ranges of baseline and multi-task models still largely overlap, the multi-task model fares significantly better when there are enough instances for each annotator head to learn from. The multi-task model also reported lower standard deviation in performance than the baseline model, suggesting better robustness in the learned model.

The main advantage of our multi-annotator model is the ability to capture multiple perspectives efficiently. In that respect, our model fared better at modeling annotator uncertainty across the board. As shown in Figure 4, our multi-annotator model obtained better correlation overall with annotator disagreement than Softmax and MC dropout approaches across all six emotions, both in the full dataset as well as the subset (*joy* in the full dataset being the only exception). This further demonstrates the strength of our approach that does not come at any significant cost of performance or efficiency (training the multi-task model on the full dataset takes 6.1 minutes per epoch, comparing to 5.2 minutes for the baseline model).

	Annotation		Multi-task		Baseline
Text	Label	%Hate	Label	%Hate	Label
Agreed. Traitors deserve to be shot. PERIOD.	Not-Hate	0.33	Not-Hate	0.39	Hate
Are the democrats guilty of treason?	Not-Hate	0.25	Not-Hate	0.39	Hate
I don't know, you tell me ya prick.	Not-Hate	0.33	Not-Hate	0.39	Hate
They should force the 'refugees' to wear this crap.	Hate	0.67	Not-Hate	0.33	Hate
At this point, nobody gives a shit!	Hate	0.67	Not-Hate	0.44	Hate
It's OK to be a Nazi!!	Hate	1.00	Not-Hate	0.44	Hate
Arise WHITE PEOPLE! We're back.	Not-Hate	0.33	Hate	0.56	Not-Hate
Shocking! A jew defending another jew! Who'd have thought?	Not-Hate	0.33	Hate	0.56	Not-Hate
I am beginning to think being straight is misconduct.	Not-Hate	0.0	Hate	0.56	Not-Hate
Armenia is a nation of mongrel bastards.	Hate	1.0	Hate	0.78	Not-Hate
Hope they both get AIDS.	Hate	1.0	Hate	0.72	Not-Hate
I am so NOT afraid of you gay boy.	Hate	0.67	Hate	0.83	Not-Hate

Table 4: Examples from the GHC, for which the baseline differ from multi-task predictions' majority vote. (We acknowledge that individual readers may disagree with the annotation labels presented above.)

5 Analysis

In this section, we further analyze the multi-task model and its outputs, as it posted the overall best performance among the three approaches, considering the predictive performance, uncertainty modeling correlation, and time efficiency. We focus on the GHC model for this analysis.

5.1 Error Analysis

We first qualitatively analyze the mismatches between the multi-task and baseline model on their majority vote predictions. Among all GHC instances ($|X| = 27,665$), the multi-task and baseline models disagreed on 1,945 labels. Table 4 shows some examples of such instances and the corresponding majority vote, and the percentage of annotators who labeled them as hate speech. Table 4 also provides the baseline model's prediction (columns 6), the multi-task model's majority label, and the percentage of prediction heads labeling them as hate speech (columns 4 and 5).

The most common type of mismatch (57.94% of mismatches) occurs when an instance deemed non-hateful (by majority vote of annotations) is correctly labeled by the multi-task model but incorrectly labeled by the baseline (first set of rows in Table 4). In other words, these samples represent the baseline model's false-positive predictions, most of which include specific tokens, such as slur words and social group tokens. The next most common type of model mismatch (22.31% of mismatches) occurred when an instance that was deemed hateful (by majority vote) is mislabeled by the multi-task model and labeled correctly by the baseline model. In general, these two types of mismatches correspond to the positive predictions of the baseline model. A possible explanation for the frequency of such mismatches is the high rate of positive predictions by the baseline model, which is also supported by the higher recall and lower precision scores of the baseline model (Table 1).

The other two types of mismatches occurred when the baseline and multi-task model respectively predicted hateful and non-hateful labels. When this mismatch is over an instance deemed hateful by majority vote of annotations (12.19% of mismatches) the multi-task model is making a false-positive error and we observe mentions of social group names in the text. A large number of such instances had even split (54%–44%) between labels across individual predictions (see Table 4), suggesting the model was unsure. The least common type of disagreement is over instances deemed as hateful by both majority vote of annotations and our multi-task model, but mis-classified by the baseline model (7.56% of mismatches).

5.2 Uncertainty vs. Error

Now, we investigate whether the uncertainty in predictions is correlated with whether the multi-task model was able to correctly predict the majority label. Note that the value of uncertainty, based on Equation 1, falls between 0 and 0.25. We observe that the mean value for uncertainty in correct predictions was 0.049 compared to 0.170 when the model was incorrect. Figure 5a shows the corresponding violin plots. While most incorrect predictions had high uncertainty, a small

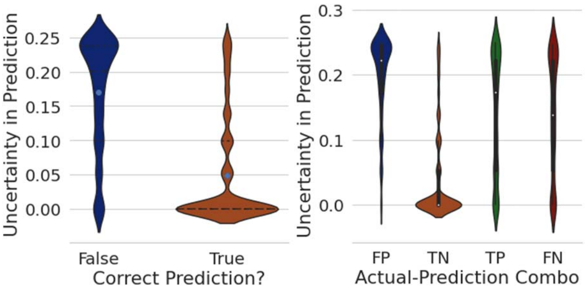

Figure 5: Violin plots denoting distribution across uncertainty for true-positive, false-positive, false-negative, and true-negative predictions on GHC.

but significant number of errors were made with certainty.

Separating this analysis across true positives, false positives, false negatives, and true negatives represents a more informative picture. For instance, the model is almost always certain about true negatives (M(uncertainty) $= 0.040$). Similarly, the model is almost always uncertain about false positives (M(uncertainty) $= 0.199$), something we also observed in the error analysis presented in Section 5.1. On the other hand, both true positives and false negatives have a bi-modal distribution of uncertainty, with similar mean uncertainty values of 0.140 and 0.141, respectively. In sum, a negative prediction with high uncertainty is more likely to be a false negative, in our case.

6 Discussion

We presented multi-annotator approaches that predict individual labels corresponding with each annotator of a subjective task, as an alternative to the more common practice of deriving (and predicting) a single ''ground-truth'' label, such as the majority vote or average of multiple annotations. We demonstrate that our method based on multi-task architecture obtains better performance for modeling each annotator (63.2 F_1-score, micro-averaged across annotators in GHC), and even when aggregating annotators' predictions, our approach matches or outperforms the baseline across seven tasks. Our study focuses on majority vote as the baseline aggregation approach to demonstrate how this commonly used approach loses meaningful information. Other aggregation strategies such as MACE (Hovy et al., 2013) and Bayesian methods (Paun et al., 2018) could be explored in future work as complementary approaches that can work with the multi-annotator framework.

6.1 Advantages of Multi-Annotator Modeling

One core advantage of our method, which can further be leveraged in practice, is its ability to provide multiple predictions for each instance. As demonstrated in Figures 2 and 4, the multiple predictions can derive an uncertainty estimation that better matches with the disagreement between annotators. The estimated uncertainty could be used to determine *when not to make a prediction* or to route the example to a manual content moderation queue as it may be an example that annotators likely disagreed on. One could also investigate how to learn an uncertainty threshold to make cleverer predictions. For instance, based on our analysis in 5.2, a negative prediction with high uncertainty is very likely to be a false negative. One could use this knowledge in a deployment scenario and predict a positive label in case of a negative majority prediction with high uncertainty.

Predicting multiple annotations rather than a ground truth is specifically essential in subjective tasks. As Alm (2008) argues, in many subjective tasks, the aim is not to find an *accurate* answer; instead, a model can produce the most *acceptable* answer based on responses from different judgements. Accordingly, our method contrasts with approaches for enhancing ground-truth generation prior to modeling. Our approach aims to preserve annotators' consistency in labeling by delaying the annotation aggregation until the final stage. As a final step, if required, application-driven approaches can be employed to find the most proper answer. For instance, an aggregation approach based on MACE (Hovy et al., 2013; Paun et al., 2018), could be applied to the predicted individual labels to find a final label that considers the trustworthiness of individual annotators.

Researchers have pointed out that in more objective tasks, such as commonsense knowledge or word sense disambiguation, training a model on judgements of a specific set of annotators lack generalizability to annotations generated by new annotators (Geva et al., 2019). However, in subjective tasks such as affect and online abuse detection, different annotator perspectives, and their contrasts, can be useful (Gordon et al., 2021).

Another advantage of having multiple prediction heads in a multi-task architecture is that we could adapt the same model to different value systems. For instance, in cases where annotators with different moral beliefs systemically produce different labels (Waseem, 2016; Díaz, 2020; Patton et al., 2019), one could use the multi-task approach to have a single global model that can adjust predictions to be conditioned on different value systems. This is valuable for international media platforms to build and deploy global models that attend to local cultures and values without retraining entirely separate models for each culture.

Multi-annotator modeling can also be applied in scenarios that may benefit from obtaining several perspectives for a single instance. For example, in detecting affect in language, a range of subjective human knowledge, interpretation, and experience can be modeled through a multi-annotator architecture. This approach would generate a range of affective states either along affect categories, such as anger and happiness, or dimensions, such as arousal and pleasantness (Alm, 2011, 2008), which correspond with different subjective perceptions of the text. Another example is sarcasm detection, where an ambiguous sarcastic text is labeled differently according to annotators' thresholds for sarcasm (Rakov and Rosenberg, 2013). In a multi-annotator setting, internal consistency of each annotators' threshold for sarcasm may be preserved in the training process.

6.2 Limitations and Challenges

Our approach is not without limitations. Our experiments were computationally viable because of the relatively small number of annotators in our annotator pool (18 for GHC and 82 for the GoEmotions dataset), which is not usually the case with large crowd-sourced datasets. For instance, the dataset by Díaz (2020) has over 1.4K individual annotators, and Jigsaw (2019) built a dataset with over 8K annotators. Fine-tuning that many separate annotator heads will be computationally expensive and may not be a viable option. However, clustering annotators based on their agreements and aggregating annotator labels into cluster labels could address this issue. In that scenario, the multi-task model would include separate classifier heads for each cluster of annotators. The number of clusters could be determined based on availability of computational resources and data factors to enhance the multi-task approach. This is an important direction of research for future work.

The proposed approach along with other methods for incorporating individual annotators and their disagreements are only viable when annotated datasets include annotator-level labels for each instance. However, most multiply annotated datasets contain only per-instance majority labels (Waseem and Hovy, 2016; Jigsaw, 2018), or aggregate percentages (Davidson et al., 2017; Jigsaw, 2019). Even in cases where the raw annotations were released, the multi-annotator model requires there being enough annotations from each annotator to model them effectively. However, we observed that the dataset designers may not have envisioned such a utility of annotator-level labels for downstream analysis. For instance, in the GoEmotions dataset, many annotators labeled fewer than 1000 instances, making it hard for annotator-level modeling. Moreover, the high cost of gathering large number of annotations per annotator in crowdsourcing platforms may limit the data collection and call for post-hoc modeling solutions. One way to tackle this issue is by choosing a subset of top-performing annotator heads (during the validation step) for the final prediction. Future work should look into such post-processing steps that could further improve the performance.

To enable further exploration into open questions in studying annotator disagreements and efficient ways to model them, the main challenge is the lack of annotator-level labels. This largely stems from the practice of considering crowd annotators as interchangeable, and not accounting for the differences in their perspectives. We recommend data providers to consider releasing individual annotation labels, when feasible to do so, in an anonymized way and with appropriate consent. We also encourage researchers to design data collection efforts in a way that includes a sufficient number of annotations by each annotator, so that systematic differences in their annotation behaviors could be better understood and accounted for.

7 Conclusion

We present a multi-annotator approach that employs a different classifier head for each annotator of a dataset as an alternate method to the practice of predicting the aggregated majority vote. We demonstrate that our method can efficiently obtain better performance in modeling each annotator as well as match the majority vote prediction performance. We present experiments across different subjective classification tasks, including hate speech detection and six different emotion detection tasks. The model uncertainty estimated based on our multi-annotator model(s)' predictions obtains a higher correlation to the annotation disagreement than more traditional methods. We expect future work to investigate our multi-annotator approach as a means to detect and mitigate model biases. Moreover, monitoring the performance of annotator heads and model uncertainty in an active learning setting has the potential to capture a more diverse and comprehensive set of perspectives in data.

8 Ethical Considerations

Our paper discusses an approach for attending to individual annotator's judgements in training a supervised model. In doing that, our multi-annotator approach better preserves minority perspectives that are usually sidelined by majority votes. Our intended use case for this approach is in subjective NLP tasks, such as identifying affect, abusive language, or hate speech, where generating a single true answer does not capture the nuances. While our method likely preserves minority perspectives, a misuse of this technique might happen upon weighting individual annotator's labels during prediction. Such an alternation aimed solely to improve the majority label prediction performance may adversely impact the representation of different perspectives in the model. In fact, such an optimization may cause further marginalization to under-represented perspectives than the current majority vote–based approaches. For instance, identifying annotator heads that significantly disagree with the majority vote might cause their perspectives to be at higher risk of being excluded.

It is also important to consider the number of annotators in the annotator pool when applying this method, in order to protect the privacy and anonymity of annotators, since our approach attempts to model their personal subjective preferences and biases. This is especially critical in the case of sensitive tasks such as hate speech annotations, where associating individual annotators with such representations may be undesirable.

Acknowledgments

We thank Ben Hutchinson, Lucas Dixon, and Jeffrey Sorensen for valuable feedback on the manuscript. We also thank TACL Action Editor, Dirk Hovy, and the anonymous reviewers for their constructive feedback.

References

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* Text and Speech*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Héctor Martínez Alonso, Anders Johannsen, Oier Lopez de Lacalle, and Eneko Agirre. 2015. Predicting word sense annotation agreement. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 89–94. `https://doi.org/10.18653/v1/W15-2711`

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer. `https://doi.org/10.1007/978-3-540-74628-7_27`

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono. 2018. Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4964–4968. IEEE. `https://doi.org/10.1109/ICASSP.2018.8461299`

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1664–1674. https://doi.org/10.18653/v1/D19-1176

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585. https://doi.org/10.18653/v1/E17-2092

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30. https://doi.org/10.1145/3359276

Veronika Cheplygina and Josien P. W. Pluim. 2018. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 105–111. Springer. https://doi.org/10.1007/978-3-030-01364-6_12

Huang-Cheng Chou and Chi-Chun Lee. 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890. IEEE.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22. https://doi.org/10.1145/3377323

Gloria Cowan and Désirée Khatchadourian. 2003. Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly*, 27(4):300–308. https://doi.org/10.1111/1471-6402.00110

Alan Cowen, Disa Sauter, Jessica L. Tracy, and Dacher Keltner. 2019. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90. https://doi.org/10.1177/1529100619850176, PubMed: 31313637

Crowdflower. 2016. https://www.figureeight.com/data/sentiment-analysis-emotion-text/

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3504

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28. https://doi.org/10.2307/2346806

Marie-Catherine De Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333. https://doi.org/10.1162/COLI_a_00097

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association* *for Computational Linguistics*, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358. https://doi.org/10.1016/j.eswa.2013.05.050

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Mark Díaz. 2020. *Biases as Values: Evaluating Algorithms in Context*. Ph.D. thesis, Northwestern University.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14. https://doi.org/10.1145/3173574.3173986

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

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200. https://doi.org/10.1080/02699939208411068

Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 566–570. IEEE. https://doi.org/10.1109/IJCNN.2016.7727250

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204

Gavin Gaffney. 2018. Pushshift gab corpus. https://files.pushshift.io/gab/. Accessed: 2019-5-23.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1107

Asma Ghandeharioun, Brian Eoff, Brendan Jou, and Rosalind Picard. 2019. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206. IEEE. https://doi.org/10.1109/ICCVW.2019.00517

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445423

Rainer Greifeneder, Herbert Bless, and Klaus Fiedler. 2017. *Social cognition: How individuals construct social reality*. Psychology Press. https://doi.org/10.4324/9781315648156

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

Julia Hirschberg, Jackson Liscombe, and Jennifer Venditti. 2003. Experiments in emotional speech. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266. `https://doi.org/10.1126/science.aaa8685`, PubMed: 26185244

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 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*, pages 1120–1130.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.487`

Jigsaw. 2018. Toxic comment classification challenge. Accessed: 2021-05-01.

Jigsaw. 2019. Unintended bias in toxicity classification. Accessed: 2021-05-01.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P19-1357`

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

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2020. The gab hate corpus: A collection of 27k posts annotated for hate speech. `https://doi.org/10.31234/osf.io/hqjxn`

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michael Kläs and Anna Maria Vollmer. 2018. Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In *International Conference on Computer Safety, Reliability, and Security*, pages 431–438. Springer. `https://doi.org/10.1007/978-3-319-99229-7_36`

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112. `https://doi.org/10.1080/19312458.2011.568376`

Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Eighth European Conference on Speech Communication and Technology*.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2(2010):627–666.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 125–132. `https://doi.org/10.1145/604045.604067`

Tong Liu. 2020. Human-in-the-loop learning from crowdsourcing and social media.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of* *the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.296

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE. https://doi.org/10.1109/ACII.2009.5349500

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer. https://doi.org/10.1007/978-3-030-36687-2_77

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 557–566. https://doi.org/10.1145/1743384.1743478

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278. Barcelona, Spain. https://doi.org/10.3115/1218955.1218990

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2311–326. https://doi.org/10.1162/tacl_a_00185, https://doi.org/10.1162/tacl_a_00204

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2019.260

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585. https://doi.org/10.1162/tacl_a_00040

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1078

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion, *Theories of Emotion*, pages 3–33. Elsevier. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the 15th Linguistic Annotation Workshop*, Virtual. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1578

Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo, and Bertie Vidgen. 2020. Online abuse and human rights: WOAH satellite session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.1

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.15

Rachel Rakov and Andrew Rosenberg. 2013. ‘‘sure, i did the right thing’’: A system for sarcasm detection in speech. In *Interspeech*, pages 842–846. https://doi.org/10.21437/Interspeech.2013-239

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. https://doi.org/10.1145/1753846.1753873

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145. https://doi.org/10.1037/0033-295X.110.1.145, PubMed: 12529060

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. https://doi.org/10.18653/v1/W17-1101

Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613751

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74. https://doi.org/10.3115/1621474.1621487

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.132

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142. https://doi.org/10.18653/v1/W16-5618

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*. https://doi.org/10.18653/v1/W17-3012

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. https://doi.org/10.18653/v1/N16-2013

Tamsyn P. Waterhouse. 2013. Pay by the bit: an information-theoretic metric for collective human judgment. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 623–638. https://doi.org/10.1145/2441776.2441846

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308. https://doi.org/10.1162/0891201041850885

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. pages 3143–3155. https://doi.org/10.18653/v1/2021.eacl-main.274

Minghao Zhu, Youzhe Song, Ge Jin, and Keyuan Jiang. 2020. Identifying personal experience tweets of medication effects using pre-trained RoBERTa language model and its updating. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 127–137, Online. Association for Computational Linguistics.